Resource Allocation and Scheduling for Communication Satellites with Advanced Transmission Antennas

by

Jihwan Patrick Choi

B.S., EE, Seoul National University (1998)

S.M., EECS, Massachusetts Institute of Technology (2000)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author .....
Department of Electrical Engineering and Computer Science
September, 2006

Certified by .....
Vincent W. S. Chan
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Resource Allocation and Scheduling for Communication Satellites with Advanced Transmission Antennas

by

Jihwan Patrick Choi

Submitted to the Department of Electrical Engineering and Computer Science
on September, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

For multimedia and other data services over satellite networks, the efficient management of scarce satellite communication resources is critical for the economic competitiveness of the medium. To support a broad spectrum of users with small terminals at high data rates, narrow transmit spotbeams from the satellite must be used. Since satellite on-board resources are too expensive to illuminate all of the spotbeam-coverage cells within the satellite service area, an optimized method of agile antenna gain patterning and beam scheduling is required to greatly improve the efficiency of transmission and power management.

In this thesis, we jointly optimize resource allocation/scheduling, congestion control and antenna gain patterning for communication satellites with advanced transmission antennas. Then, we develop a low-complexity on-line algorithm that considers channel conditions, interference and average delay constraints, and approaches the theoretical steady-state limit.

We introduce optimized beam profiling based on traffic demand and channel conditions over satellite downlinks, which can achieve a substantial power gain and reasonable proportional fairness. We show that a modest number of active parallel beams are sufficient to cover many cells efficiently with dynamic capacity allocation. Next, for the multiple beam antenna case, we develop a jointly optimized scheme of beam allocation and congestion control with transmitter-sharing and average delay constraints, which provides high throughput and/or small average queueing delays. Last, we find the solution for joint antenna gain patterning and scheduling by considering spatially close co-channel interference in the use of phased array antenna. We suggest an optimum scheduling policy, which selects users with higher marginal returns of a composite cost function with respect to allocated power, in terms of better channel conditions, less interference (depending on users' geographic distribution), and larger delay. The simulation result indicates that a real-time on-line algorithm can achieve a throughput close to the analytic steady-state upper bound. Due to its flexible power allocation, we demonstrate that the phased array antenna can provide

better performance than the multiple beam antenna when a small number of users are very demanding or many users are located in a small and crowded area.

Thesis Supervisor: Vincent W. S. Chan
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my advisor, Professor Vincent Chan for his encouragement, patience, support and guidance for the years. This thesis has benefited greatly from his deep and broad knowledge and keen engineering intuition. His emphasis on logical and creative thinking, precise communication/presentation skills and strong leadership for "a real leader and synthesizer" was priceless to my professional development.

I would like to thank my thesis committee members, Professor Joel Schindall and Professor Asuman Ozdaglar for their helpful feedbacks and suggestions.

I would like to thank DARPA and NRO for their financial support to this research. I am also grateful to the KFAS (the Korea Foundation for Advanced Studies) for supporting my study at MIT.

My officemates have made my life at 02139 much more enjoyable and meaningful. Thanks to Etty, Guy, Kyle, Lillian, Nick, Roop, Dr. Serena and Yonggang for invaluable discussions and fun-filled chatting.

My Mother and Father have given me all they could and been always my co-No. 1 fans. I would not stand where I stand now without their unconditional trust and love. I give my deep appreciation to my parents.

Yoonjung has been not only my Lady but the best friend throughout ebbs and flows of my MIT years. Last but not least, I thank Yoonjung for her encouragement, patience, prayer and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Satellites are an indispensable communication medium in many areas. People watch live events that happen on the other part of the globe. Satellite phones continue to provide voice and data (primitive) communications almost all around the world. Satellite communications offer an economical alternative to terrestrial media. In the US, many suburban and rural residents are subscribing DirecTV or EchoStar and one or two-way HughesNet (recently renamed from DirecWay) for their TV and Internet access. Recently, satellite radios of XM and Sirius have been a huge success. Some countries, such as Canada that has a huge area with sparse population and Indonesia that consists of a lot of islands, have a steady demand for new communication, but deploying new fiber or wireless infra may not be economically viable.

A key advantage of satellite communications is its ability to provide services for mobile users as well as fixed ground users in sparsely populated areas, over oceans, and in the air, where terrestrial infrastructure is not available or too expensive to deploy. Traditionally, satellites provide trunk and back-up routes for terrestrial networks. Satellites can broadcast information to a large area, such as an entire country, and there are examples of successful business providing satellite TV and radio broadcasting. Data networking as an access to the Internet is a new application with a promising future for satellite systems.

The satellite-to-Earth channel has long link distance with a long propagation

delay (especially for geostationary orbit, GEO, satellites), time-varying link quality due to weather, and sharing by multiple users in the form of multiple access and broadcasting. GEO satellites at the altitude of 40,000 km has a round-trip delay of approximately 250 msec. A time-varying link quality due to weather is appreciable especially in the high frequency band (e.g. 20 GHz) because of absorption by moisture. The unique properties of satellite channels dictate the need for new network protocols, unlike those for terrestrial media. Satellite systems will have many users within their coverage areas, but only have a limited amount of precious on-board resources, such as power, transmitters, amplifiers, and receivers.

This thesis studies the optimum resource allocation and scheduling problem for an advanced data communication satellite system. In this chapter, we give some motivations for the problem, a review of related work, our contributions and an outline of the thesis.

1.1 Motivation

Traditionally, satellite communications have been used for the applications of telephony and data trunking primarily based on circuit switching, and broadcasting of video and audio. The demand for Internet connection will extend to mobile and fixed users without terrestrial infrastructure support for cost or time to deployment reasons. Thus, in the future, a significant focus of the satellite industry may shift to broadband packet data networking over satellites. It is vital to design satellite network architectures to be cost-competitive with respect to the terrestrial media of wireless and fiber. The traditional architecture designed for long duration circuit traffic will not be appropriate to meet the bursty traffic requirements of future data networks over satellites. A new packet (datagram) based design is essential. In this thesis, we first examine the theoretical steady-state limit by solving static optimization problems, and identify the operation points of current systems. Then, we develop dynamic algorithms to approach the steady-state upper bound.

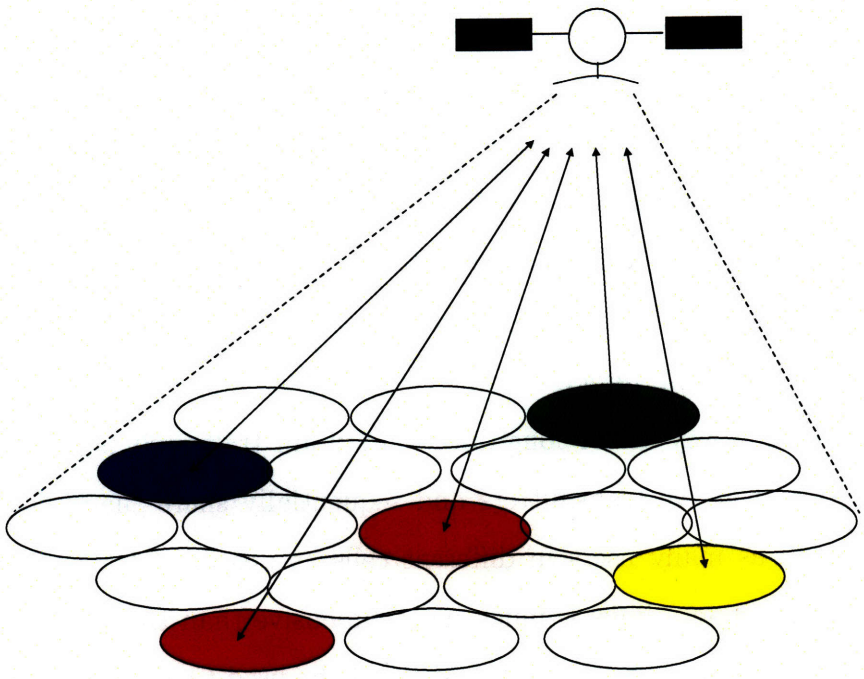

Figure 1-1: A satellite with multiple narrow spotbeams

A future satellite will use many narrow spotbeams in its coverage area (Fig. 1-1). Narrow spotbeams can project a higher power density than broad beams and thus can support higher data rates to small user terminals. In addition, the same frequencies can be reused in different cells, increasing the total system capacity. As higher frequency bands (20 GHz and beyond) are used to provide higher rates for data networking applications, it becomes more attractive architecturally to implement rapidly reconfigurable, agile and narrow beams, so precious resources are used efficiently. However, with narrow spotbeams, a large number of beams and transponders would be required to cover the service area. For example, consider a GEO satellite at an altitude of $L = 22,000$ miles. Assume that the satellite has a transmit antenna with a diameter of $D = 10$ m at 20 GHz with carrier wavelength $\lambda = 0.015$ m. The diffraction-limited narrowest spotbeam size is then

$$\frac{\lambda L}{D} = 33 \text{ miles.} \tag{1.1}$$

To cover the whole United States ($\sim$ 3,000 x 2,000 square miles), the satellite would need to generate 5,400 beams and will be impossible to carry corresponding transponders and on-board equipments. Even with a smaller antenna size of $D = 3$ m, the beamwidth is 110 miles and the satellite still needs about 600 beams. On-board resources such as power, modulators, amplifiers and receivers are expensive and consume considerable weight and power. In addition, data application users have bursty traffic (inactive for a long period of time $\sim$ 99 %). It is inefficient if resource is allocated in the same way as for long-duration stream traffic. It is possible to have a small number of active transmitters and sequentially share them efficiently, and equitably among as many users in different cells as possible. Capacity over satellite RF (radio frequency) links is much more expensive than that over fibers due to hardware cost. Dynamic capacity allocation must be used to achieve high utility and cost-competitive satellite communication systems.

The issue of providing fair and efficient resource allocation and channel access should be considered together with system performance guarantees on throughput and delay. In future broadband packet data networks, packetized traffic with different user needs (e.g., amount of demand, priority of connection, quality of service, price, etc.), which are time-varying and location-dependent, should be given service differentiation while achieving high throughput and revenue maximization simultaneously. In satellite networks, due to open air channel conditions, connectivities and capacities of up/downlinks are changing dynamically and randomly (sometimes at a time scale of seconds [8, 9]). High frequency bands are very vulnerable to signal attenuation due to bad weather, especially rain. Thus, resource sharing and scheduling for multibeam satellite downlink transmission must be adaptive to nonuniform traffic and fading channel states, while meeting delay constraints for real-time applications, if the satellite system is to be operated efficiently.

A good satellite network should incorporate an efficient architecture for data routing and congestion control. Congestion control is critical to avoid excessive packet loss in practical systems with finite-size buffers while providing an acceptable queueing

delay. By throttling incoming traffic rates, one can stabilize the system and allocate resources efficiently with some degree of fairness.

The most challenging design task is to maximize efficiency by considering the problem from the viewpoint of joint optimization over multiple network layers. These crosslayer issues include

1. high dynamic range of adaptive rate transmission and power allocation over frequency bands with high signal attenuation in the Physical Layer,

2. beam scheduling for broadcasting and multiple access of bursty users in the MAC (medium access control) Layer,

3. possibility of point-to-point ARQ (automatic repetition request) for reliability in the DLC (data link control) Layer,

4. routing amongst satellites and ground stations over random-changing physical channels in the Network Layer,

5. efficient end-to-end flow control in the presence of channel error and congestion.

It is still an open problem to solve all these issues together because of the complexity of the problem. Thus, we may only explore some of these crosslayer problems separately and develop a feel for how an efficient system must be designed. We present in this thesis a simplified formulation of the congestion control and the satellite resource allocation problem, and explore jointly optimized solutions, in order to focus on some aspects and obtain analytical results.

Before suggesting a specific implementation, we will determine what the theoretical limit is with the ideal 'genie-aided' antenna system and protocol. At the same time, we will survey the currently deployed and proposed satellite network systems, so that they can be used as benchmarks when we assess potential improvement of future systems.

Diffraction theory [26] (also see Friis transmission equation [46]) gives a linear equation relating transmit and receive power on the far-field of satellite-to-earth link

as

$$P^r = \frac{A^t A^r}{\lambda^2 L^2} P^t, \tag{1.2}$$

where P^r is the received power, $A^t = \pi D^2/4$ is the area of the transmit antenna (D is the diameter of the transmit antenna), $A^r = \pi \delta^2/4$ is the area of the receive antenna (δ is the diameter of the receive antenna), λ is the wavelength used, L is the altitude of the satellite, and P^t is the transmit power.

When we assume no interference between K parallel beams and uniform power allocation of $P^t = P_{total}/K$ with total transmit power P_{total}, the maximum bandwidth efficiency from the band limited Shannon capacity C on a satellite-to-earth path is

$$C/W = K \log_2 \left(1 + \frac{A^r A^t}{\lambda^2 L^2} \frac{P_{total}/K}{N_0 W}\right) = K \log_2 \left(1 + \frac{\pi^2 D^2 \delta^2}{16 \lambda^2 L^2} \frac{P_{total}/K}{N_0 W}\right), \tag{1.3}$$

where N_0 is the noise power density ($= kT$ with $k = 1.38 \cdot 10^{-23} = -228.6$ dBW/HzK and $T \sim 290$ K) and W is the bandwidth used.

To cover the area of a footprint diameter $F_{coverage}$ with the narrowest spotbeam width $\frac{\lambda L}{D}$ given as in Eq. (1.1), the required number of multiple beams is

$$K_2 = \frac{\frac{\pi}{4}(F_{coverage})^2}{\frac{\pi}{4}(\frac{\lambda}{D} L)^2} = \left(\frac{D F_{coverage}}{\lambda L}\right)^2. \tag{1.4}$$

In Table 1.1, we list this theoretical maximum number K_2 together with the actual or designed number K_1, which the satellite operators claim that they are using or planning. Using the channel parameters [39] of the satellite systems[1] in Table 1.1, Table 1.2 shows the maximum bandwidth efficiency (bits/sec/Hz) of the satellite networks, based on Eq. (1.3). We observe that K_1 is quite a bit smaller than K_2 for all the systems. In fact, the current systems deploy a small number of spotbeams by time-sharing and/or enlarging some of them.

[1]Brief overviews of these four systems are given in Section 2.1. Globalstar and Iridium are operating currently with the parameters listed here. The parameters of Teledesic and ICO in Table 1.1 are given as of 2001. Since then, two have merged and changed their plan to the GEO system of "new ICO."

Networks	D	δ	λ	L
ICO	2 m	10 cm	0.15 m (2 GHz)	10,390 km
Iridium	30 cm	15 cm	0.2 m (1.6 GHz)	780 km
Globalstar	30 cm	20 cm	0.12 m (2.5 GHz)	1,414 km
Teledesic	12 m	30 cm	0.015 m (20 GHz)	1,375 km

Networks	P_{total}	W	$F_{coverage}$	K_1	K_2
ICO	500 W	30 MHz	12,900 km	163	274
Iridium	400 W	16.5 MHz	4,650 km	48	80
Globalstar	380 W	16.5 MHz	5,760 km	16	104
Teledesic	2000 W	500 MHz	N/A	N/A	N/A

Table 1.1: Channel parameters of satellite network systems

Networks	Bandwidth efficiency in systems (bits/sec/Hz)	Shannon limit of bandwidth efficiency with K beams (bits/sec/Hz)			
		$K = 1$	$K = K_1$	$K = K_2$	$K = \infty$
ICO	0.96	5.5	54.2	56.7	61.0
Iridium	1.2	8.3	139.3	183.0	448.1
Globalstar	0.73	8.8	77.5	249.3	640.4
Teledesic	10	24.2	N/A	N/A	$2.71 \cdot 10^7$

Table 1.2: Bandwidth efficiency comparison of satellite network systems

The numbers in the column of "Bandwidth efficiency in systems for uniform traffic" are from the following calculations. We assume uniform traffic distribution and no battery energy limitation,[2] so the numbers are idealistic.

1. ICO: maximum 38.4 kbps per TDMA carrier · 750 TDMA carriers per satellite / 30 MHz for uplink = 0.96

2. Iridium: 50 kbps per 10 channels / (31.5 kHz per 10 channels + 10.2 kHz for guard band per 10 channels) = 1.2

3. Globalstar: maximum 9.6 kbps per channel · 2500 total channels per satellite / (16.5 MHz · 2) for uplink and downlink = 0.73

4. Teledesic: 10 Gbps per satellite / (0.5 GHz · 2) for uplink and downlink = 10

We can see that the current system efficiencies are far less than the theoretical Shannon limits, as much as 20 dB away when we compare the bandwidth efficiency in systems with the Shannon limit at $K = K_1$. This is partially because our calculation of Shannon limits is based on such idealized assumptions with no interference, no rain attenuation, uniform traffic, the full service of the entire coverage area, and operation at the Shannon limit. Nevertheless, the current systems are operating very far from the limit of bandwidth efficiency ($\simeq$ 1 bits/sec/Hz, which corresponds to that of BPSK, binary phase shift keying) and there is plenty of room for improving efficiency. System capacity increases significantly with multiple spotbeams rather than with a single scanning beam. However, in practice, illuminating all area all the time causes waste of resource due to geographical and temporal fluctuation of real traffic. If the beams are narrow and there are a large number of beam sizes in the coverage area, it is also impossible to have as many transmitters and receivers in the satellite. This motivates resource (such as transmitters and receivers) allocation and

[2]Especially in the design of low-earth-orbit (LEO) satellites such as Globalstar and Iridium, the problem of battery energy charge/discharge and allocation is as important as that of the bandwidth efficiency. The satellites should be equipped with efficient solar panels and thermal heat dissipation schemes to be cost-effective.

scheduling based on traffic demand, channel conditions, delay deadlines and spatially close co-channel interference.

1.2 Related Work

Even though the multilayer optimization problem, especially specific to the satellite medium, remains unsolved, there have been prior works on the architecture of each separate network layer. Some examples for multibeam satellite systems include the design of antennas [16, 17, 48] and multiple access protocol [37]. In a US patent by Black *et al.* [5], for multimedia application, an on-board switching system using a specific packet structure and an up/downlink beam management technique are suggested to route and schedule data packets from the source terminal to the destination. The proposed system is adaptive to traffic requirements and includes countermeasures for rain attenuation. However, the description in [5] does not provide systematic problem modeling and solutions. This thesis will emphasize the mathematical formulation and analytic solutions of the optimization problem.

Congestion control problems over satellite networks have been addressed by modifying the existing Transport Layer protocols or suggesting new protocols, to overcome the disadvantages of long propagation delay or bursty errors over satellite channels. Some examples are Satellite Transport Protocol (STP) [29], TCP-Peach [1], and eXplicit Control Protocol (XCP) [34]. Admission control for packet or circuit connection in satellite communication systems is used for reducing system congestion and managing network resources efficiently. In Jamalipour and Ogawa's work [32], traffic distribution and distances from users to satellites are considered as important factors to control admissions and transmissions of packets, in order to mitigate the effect of multiple access interference on the throughput degradation in a low earth orbit (LEO) satellite network. In Koraitim and Tohme's work [38], a dynamic threshold algorithm is proposed for resource sharing and admission control at the MAC Layer in multi-class service satellite networks, to provide good channel utilization and to

reduce the bursty data delay. In Siwko and Rubin's work [51], for capacity-varying networks, the information on stochastic changes of future capacity is used to provide a control of trading-off between admission blocking and connection dropping. A part of this thesis develops a joint optimum policy of congestion control and beam switching under average delay constraints.

There are several precedents on the problem of power and server allocation in multibeam satellite systems. In Neely *et al.*'s work [44], a power allocation policy is suggested to stabilize the system as long as the arrival rate vector is inside the capacity region, and is based on the amount of unfinished work in the queue and the channel state, without the knowledge of traffic arrival rates. When the users are covered by multiple satellites, each of which has multiple queues for downlink traffic, a routing decision is made for the maximum total throughput. In Neely and Modiano's work [45], the strategy is extended to incorporate flow control for general networks when the arrival rate is outside the capacity region. The flow control algorithm is performed based only on the current backlog, not on other parameters of arrival rates, channel states or network topology. In Ganti's thesis [23], to minimize the expected total queue length, it is shown to be optimal to serve the K longest connected queues in i.i.d. (independent and identically distributed) symmetric on-off channel conditions. In this model, when the channel is on, each channel has an identical channel condition and transmits the same maximum amount of packets. In Fu *et al.*'s work [20, 21], an energy allocation problem is studied to maximize expected data throughput or to minimize energy, subject to sending a limited amount of data over a time-varying satellite channel. The optimal scheduling policies are obtained from dynamic programming formulations, but only a single server is considered. There is still a gap between the solutions in the special cases and the general methodology that is needed for practical system implementation, such as packet delay requirement and beamforming/antenna gain patterning under co-channel interference.

1.3 Contributions

In this thesis, we define the steady-state performance limit of the best crosslayer architecture design by solving a simple joint optimization problem of resource allocation/scheduling and congestion control. We develop practical and near-optimum on-line algorithm by taking into account for the impact of different types of antenna technology, addressing the impact of different objectives for system optimization, and incorporating the very important delay criterion. In the phased array antenna case, we consider spatially close co-channel interference and include antenna gain patterning for the joint optimization problem.

In Chapter 3, we obtain the optimum solution for satellite downlink multibeam power allocation based on traffic demand and link qualities. We suggest and compare different cost functions for power and beam allocation, in order to provide insight in the trade-offs between maximizing total capacity and providing proportional fairness. Substantial power gains and fairness advantages can be realized by optimum power allocation for parallel multibeams. We model a practical situation, where the numbers of active beams and corresponding modulators are less than that of the cells in the coverage area. An optimum greedy solution (with delay issues suppressed) is given in terms of accumulated traffic and channel conditions.

In Chapter 4, for a satellite equipped with a multiple beam antenna, we develop a jointly optimized scheme of multibeam allocation and congestion control under beam-sharing and average delay constraints. Unless one cell has a overwhelmingly dominant demand, the joint scheme picks the congestion control parameter, which gives the system throughput with fairness among users, and the corresponding beam allocation in terms of the average of expected values of user parameters, such as incoming traffic rate, transmission rate and delay deadline. Numerical examples show that the joint scheme can outperform uniform beam allocation by providing substantially higher throughput and/or smaller average queueing delays.

In Chapter 5, we find the solution for joint antenna gain patterning and scheduling by considering spatially close co-channel interference in the use of phased array

antenna. When users are located far enough, the optimum scheme is to provide the narrowest spotbeams for the non-interfering active users. When potential interference can be significant between close-in users, the optimum pattern, which depends on the distances between users and the signal-to-noise ratio (SNR), can be one of the following: complete cancellation of interference, optimum suppression of interference, and the sequential service of close-in users. We suggest an optimum scheduling policy, which selects users with higher marginal returns of a composite cost function with respect to allocated power, in terms of better channel conditions, less interference (depending on users' geographic distribution), and larger delay. From the optimum analytic result, we develop a near-optimum, low-complexity and real-time on-line algorithm of performing active user selection, antenna gain patterning, power allocation, and admission control. The simulation result indicates that the real-time algorithm can achieve a throughput close to the analytic steady-state upper bound (up to 94% of the steady-state solution with random traffic). We show that the phased array antenna that is more flexible for power allocation and beam shaping can provide better performance than the multiple beam antenna especially when there are a small number of very demanding users or active users closer than a diffraction-limited beamwidth.

1.4 Outline

In Chapter 2, we provide background on communication satellites. Several examples of commercial communication satellites are given, and their applications and characteristics are briefly discussed. We present the current communication payload technologies and remark on the models and assumptions of the technologies that are studied in the thesis.

In Chapter 3, we examine the design of power and beam allocation over satellite downlinks in terms of Shannon capacity, jointly based on traffic distribution and channel conditions, maximizing system performance as well as achieving reason-

able fairness amongst users. First, we formulate the downlink multibeam capacity optimization problem and motivate the need for a satellite-to-earth communication strategy using parallel multibeams. There are two types of fading events in a high frequency (e.g., 20 GHz) satellite channel. The first is attenuation due to water in the form of fog, rain and snow, and the other is atmospheric turbulence. Here, we mainly consider rain attenuation, which is a slow fading event, and assume uniform attenuation across each spotbeam. This assumption is not perfect, but as we consider narrow spotbeams for higher data rates (the diameter of a beam can be about $50 \sim 100$ miles in the future), this model reflects the realistic situation better. If we use the worst attenuation within a narrow spotbeam in our analysis, it will yield a conservative bound for the performance, and in practice, that is what the satellite would use short of individual measurement to each user. There can be interbeam interference from the sidelobes of adjacent beams, to degenerate Shannon capacity. However, in Chapter 3 and 4 we consider a cellular system where cells on the Earth are covered by spotbeams and assume that interference is negligible because we generate very narrow beams over a large number of cells. In practice, we do not allow adjacent active beams and can locate active downlink beams far enough at each timeslot. The problem of co-channel interference between close-in users is discussed in detail in Chapter 5.

Then, for this simple model, we find an optimized design by minimizing a general function of the deficit between capacities allocated across cells and accumulated traffic demand. We focus on the problem of power, rate and beam allocation in the Physical Layer when the aggregate demand exceeds the total capacity from available power. We make the assumption that ultimately the Transport Layer protocol will serve the backed-off excess demands. The joint solution of resource allocation and congestion control under average delay constraints are presented in Chapter 4 and 5. Using different cost functions, we examine the corresponding trade-off between fairness and total capacity, and illustrate the advantage of optimum power allocation for parallel multibeams in terms of a power gain. We then include average transmission delay constraints and arrive at a modified steady-state solution. As the satellite uses

increasingly smaller spotbeams within the coverage area, it is important to economically time-share a small number of active spotbeams and transmitters in an optimized fashion. We discuss the practical impact of a limited number of active beams,[3] which is the need for downlink multibeam scheduling.

We consider two types of transmission antennas: multiple beam antenna in Chapter 4 and phased array antenna in Chapter 5. Their different implementations result in different power constraints and thus different performances. In Chapter 4, we couple a multibeam allocation problem with congestion control of incoming traffic over average delay constraints, assuming the use of multiple beam antenna (Fig. 1-2(a)) with traveling wave tube amplifiers (TWTA). Each multiple beam antenna feed is fed by its own TWTA, which results in a power constraint for each beam. Since it is assumed that TWTAs are driven well into saturation for efficiency with a single carrier, we can fix power at the maximum possible level when a TWTA is operating. The channel condition is quasi-static during the interval of interest due to weather-induced slow fading, and the channel capacity for each beam is considered to be constant. We assume that beam switching is very fast with no additional cost (which is idealization that can be approached with advanced electronic or electro-optical beam switching technologies [19, 42]). We formulate the efficiency maximization problem of satellite resources by considering incoming traffic with stability, deadline, and transmitter-sharing constraints. A closed-form analytical solution for joint beam allocation and congestion control is obtained by using queueing theory. We compare this jointly optimized scheme with the uniform beam allocation scheme with respect to throughput, average queueing delay, and fairness.

In Chapter 5, we consider the use of phased array antenna. A phased array antenna (Fig. 1-2(b)) uses solid state power amplifiers (SSPA) and can linearly superimpose signals at array elements by controlling an antenna-patterning matrix. Signal power can be divided among multiple channels up to the total power of the

[3]Throughout the thesis, the terminology of "beam" is used in the same meaning as "signal" or "carrier." In Chapter 3 and 4, beams are assumed to be delivered to some of cells at each timeslot. In Chapter 5 the phased array antenna transmission does not have to be constrained to a cellular system and the terminology of "pattern" is used instead of "beam."

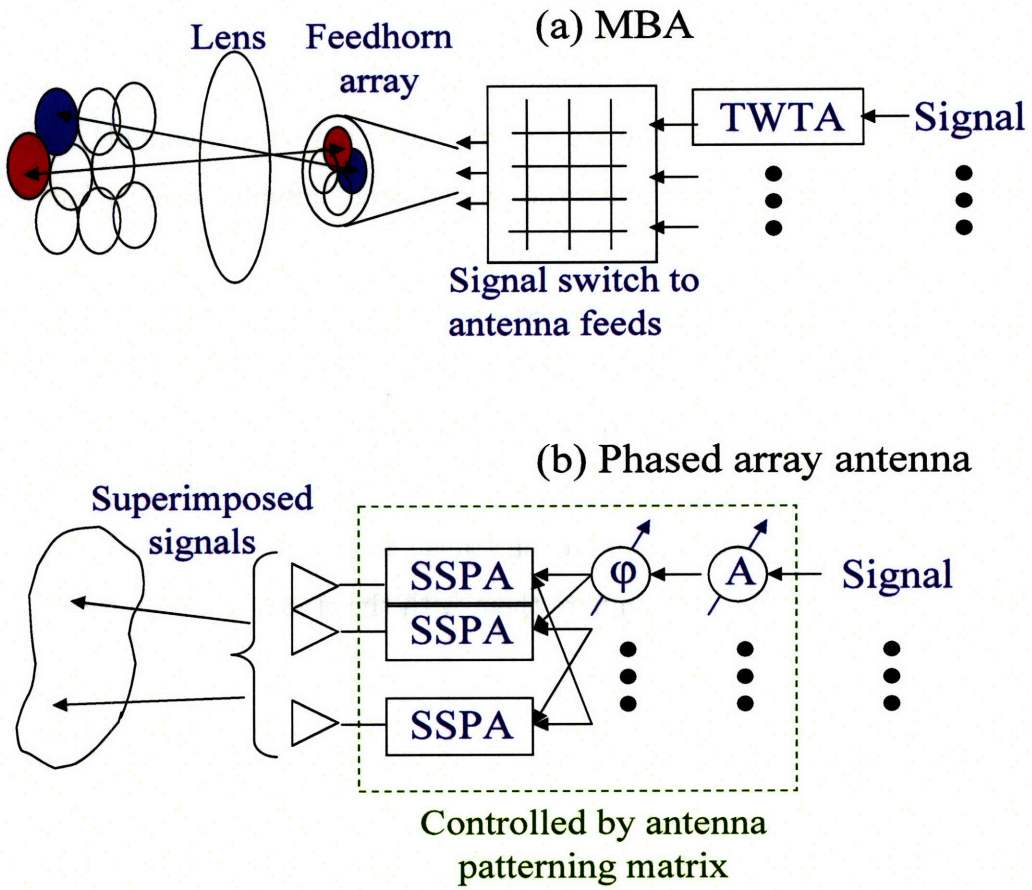

Figure 1-2: Schematics of (a) multiple beam antenna and (b) phased array antenna

array. The time-varying channel capacity is not full-on or off any more since the way of allocating power is flexible. In addition, while the multiple beam antenna has fixed beam size due to the fixed size of feedhorn for each signal, the phased array antenna can have any size and/or shape of antenna pattern by feeding many array elements with the same signal. Moreover, the phased array antenna together with transmission scheduling can be cycled much more rapidly ($\ll$ msec) than the multiple beam antenna and is advantageous in meeting time deadlines via fast switching techniques. We provide the optimum design of antenna gain patterning and scheduling of the phased array antenna adaptive to traffic distribution and channel conditions, in order to enhance efficiency. When the phased array antenna satellite has close-in active users, spatially close co-channel interference cannot be ignored any more, so that one of several interference suppression schemes from the satellite transmission antenna is deployed, depending on user-location distribution and operating signal-to-noise ratios. We compare the steady-state performance of the phased array antenna with that of the multiple beam antenna. An efficient resource scheduling algorithm is suggested for the phased array antenna transmission. We give simulation results for several types of traffic and compare them with the steady-state solution.

Chapter 2

Background

Since Arthur K. Clarke proposed the idea of "artificial satellites [11]" in 1945, communication satellites have made a drastic evolution: huge expansion of applications and innovations of communication payload technologies. In this chapter,[1] we provide some examples of communication satellites for the last half century,[2] focusing mainly on the US commercial satellites. Then, we give an overview of existent communication payload technologies[3] and remark on models and assumptions for the technologies discussed in the thesis.

2.1 Examples of Commercial Communication Satellites

2.1.1 First-generation Commercial Satellites

The first-generation commercial satellites were mainly used for telephone trunking and TV transmission between continents. The ground stations were equipped with big antennas of 15 $\sim$ 30 m diameter. One example of the first-generation satellites

[1] Most contents of this chapter will appear in the encyclopedia by UNESCO [15]. It is noted that only the parts written by the author of this thesis are presented here.

[2] Several books of [39, 41] are referred for more detail.

[3] Some textbooks of [28, 39, 40, 47, 55] are referred to for more detail.

is the Intelsat. In 1964, the International Telecommunications Satellite Consortium (Intelsat) was formed when 11 nations signed a joint agreement to design and maintain global communication satellites for commercial purposes. In 1965, the world's first commercial communication satellite, Intelsat I (also known as "Early Bird") was launched into the geosynchronous orbit. Intelsat mainly carried voice circuits, and transmitted to fixed and transportable terminals. Intelsat started from a single satellite that supported transatlantic relays, and has evolved to provide almost universal coverage, to form a global communication network. Current applications include voice and data communications, enterprise networking, financial transactions, Internet linkages, and satellite video transmission and distribution [31].

In Europe, similar to Intelsat, Eutelsat was set up as an intergovernmental organization, and launched its first satellite in 1983. It is now operating 23 satellites to provide radio and TV broadcasting services, professional data network solutions and broadband Internet access [18].

2.1.2 Second-generation Commercial Satellites

As the satellite capacity increased and ground terminals became smaller, the second-generation commercial satellites could begin to support mobile users, such as ships, aircrafts and land vehicles. The Inmarsat is an example of the second generation commercial satellites. In 1976, the Marisat system began communication services between ships and shore stations. It was combined with European Marces Satellites, to form an intergovernmental organization, and named the Inmarsat, the International Maritime Satellite Organization, in 1979.

The primary goal of the Inmarsat system was to provide global safety and maritime satellite access to mobile terminals in the ships over the ocean that terrestrial wireless infrastructure could not cover. It gradually expanded its service to users in land vehicles and aircrafts. Beginning with the satellites of Marisat, Marces and leased Intelsat V, the Inmarsat system launched Inmarsat II satellites for an increasing number of services to ships as well as airplanes in 1990. Inmarsat III satellites,

launched in 1996, could increase the capacity by the use of L band spotbeams reconfigurable for desired coverage. The Inmarsat is now operating 11 geostationary orbit (GEO) satellites and provides the worldwide service of telephony, fax and data communications [30]. The OmniTracs and the Geostar in the US and EutelTracs in Europe also provided regional services to mobile terminals in late 1980s.

2.1.3 Third-generation Commercial Satellites and beyond

Since 1990s, a significant focus of the satellite communication industry has been to support mobile communications for small hand-held terminals. For this purpose, low earth orbit (LEO) and medium earth orbit (MEO) satellite network systems have been designed and deployed because the lower altitude could give higher power density and shorter propagation delays to small ground terminals than GEO systems. However, to cover the whole globe, a larger number of satellites are needed and the coordination issues such as intersatellite routing and handover arise. Thus, a different choice of operating orbits in the third-generation commercial satellites leads to different design requirements and corresponding performances.

There are two main objectives of the third generation commercial satellites.

- To support mobile voice communications

  As fore-mentioned, satellite communications can support mobile users as well as fixed terminals, filling coverage gaps left by terrestrial infra structures, such as deserts, forests, oceans and air. Iridium and Globalstar utilize LEO satellites for mobile voice telephone service (and some limited data communications).

- To provide data services

  With a big increase of the Internet access demands, data networking over satellites can complement the terrestrial media. The Internet access via satellites is an economically viable way to connect to the Internet in suburban and rural areas where the digital subscriber lines (DSL) and cable Internet service cannot reach. The HughesNet satellite Internet system uses Ku-band while the recent

WildBlue uses Ka-band. In addition, the satellite can serve as an alternative to under-ocean fiber for data trunking and a backbone to interconnect local area networks (LAN) and metropolitan area networks (MAN).

The examples of the third-generation commercial satellite systems are briefly described in the following.

Iridium consists of 66 LEO satellites at the altitude of approximately 780 km. The system was originally designed to have 77 satellites, and was named for the element iridium, atomic number 77. Iridium provides global voice and data communications with handheld devices. Iridium LLC began service in May 1998 and declared bankruptcy in August 1999. Iridium Satellite LLC acquired all operating assets of Iridium LLC in March 2001 and resumed commercial service. Iridium has intersatellite links, so that every satellite can communicate with neighbor satellites if necessary.

Globalstar has 48 LEO satellites to cover the whole globe (except the polar regions) at the altitude of 1,414 km. It provides global telephone and data communications. Globalstar began service in October 1999 but filed for bankruptcy protection in February 2002. Different from Iridium, Globalstar does not have intersatellite links, so that the satellite is connected to the public switched telephone network (PSTN) via the ground gateway station. Since only one satellite delivers the signal between two locations on the ground (e.g., a user and a ground gateway), Globalstar is called a bent-pipe system. Approximately 60 gateways were proposed and 25 gateways were in service as of April 2002.

There are regional satellites that provide voice telephone service within some regions. Asia Cellular Satellite System began service in September 2000 and serves 24 countries in Southeast Asia. Thuraya began service in April 2001 and intends to serve 99 countries in Europe, North and Central Africa, the Middle East, Central Asia and the Indian Subcontinent.

Teledesic was proposed in 1990s for a global broadband data satellite network with a total of 288 satellites at the altitude of 1,375 km. The primary goal was to provide

high-speed Internet access over Ka band with a small propagation delay of LEO satellites. Due to the overall financial struggle of the satellite industry and decrease of market expectation, the initial plan was scaled down gradually. Teledesic changed the design of its constellation to 30 MEO satellites, after acquiring the ICO system, which was to provide global mobile personal communications services by satellite, but filed for bankruptcy protection in August 1999. However, the merge plan did not work out and the deployment of Teledesic was finally shut down in October, 2002. The new ICO instead plans S-band mobile satellite services via a GEO satellite.

Spaceway was planned as a Ka band GEO satellite system for broadband communications. After some changes of the original plan, Spaceway is expected to provide the service for high definition TV over DirecTV and two-way Internet access over HughesNet. Similar to Spaceway, Astrolink was planned as a global broadband Ka band system with nine GEO satellites, but later cancelled due to financial problems.

Another recent application of commercial satellite systems is satellite radio such as XM and Sirius in the US and WorldSpace in Europe, Africa, and Asia. Satellite radio broadcasts digital radio programs from satellites to the subscribers that have adequate radio receivers. Because of a wider coverage of satellites than terrestrial radio stations, users can receive the same signal in any place within the coverage area. XM has two fixed-location GEO satellites and Sirius has three GEO satellites that pass over North and South America. To overcome obstructions, such as buildings, of the line of sight between satellites and receivers and increase system capacity, ground repeaters are deployed to relay signals from satellites, especially in metropolitan areas.

2.2 Communication Payload Technologies

Individual satellites receive signals from ground stations/users or neighbor satellites, repeat and/or regenerate the received signals and then send out to the Earth or other satellites. In this section, first we describe the primary technologies for communication payloads: repeater/transponder, power amplifier, coding and modulation.

Since a satellite covers many users on its coverage area, there needs to be some form of coordination for users to share satellite resource efficiently. In practice, all communication satellites are operated as cellular systems by the use of multiple spotbeams. In a cellular satellite system, a coverage area is divided into small cells and each cell is covered by a spotbeam. Within each spotbeam, a multiple access scheme is used by dividing one common uplink channel into small orthogonal sub-channels, so that each user is assigned to different sub-channels, such as timeslots, frequency subbands and codes, and does not interfere with each other. Here, we explain multiple access schemes and the cellular satellite system.

Finally, as the satellite industry gets interested in the increasing demand for Internet connection and applications, satellite communication systems need to meet the request for packet data networks. The most dominant Internet protocol TCP/IP has been designed for the terrestrial fiber channel, which is usually more reliable and has a shorter delay than the satellite channel, and thus performs poorly over satellite networks. Here, we describe the basic function of TCP/IP and its modification for satellite systems. In addition, as some LEO and MEO satellite network systems are deployed for the data networking, several issues must be addressed: communications and data routing over inter-satellite links (ISL), and handover of spotbeams and satellites.

2.2.1 Transparent and Regenerative Repeaters

A satellite repeater processes the signal fed from the receiver antenna before the transmit antenna sends the signal out. A repeater converts the signal frequency into the transmit carrier frequency, and amplifies the signal power by using high power amplifiers such as traveling wave tube amplifiers (TWTA) or solid state power amplifiers (SSPA). A repeater consists of one or more channels, which is called a transponder.[4] Each transponder carries a different signal, so that multiple signals

[4]In some texts, a "transponder" is used only for a transparent repeater as a "transponder" satellite while a "processing" satellite is used for a regenerative repeater. Here, we follow the definition of transponder in Maral and Bousquet's textbook [41].

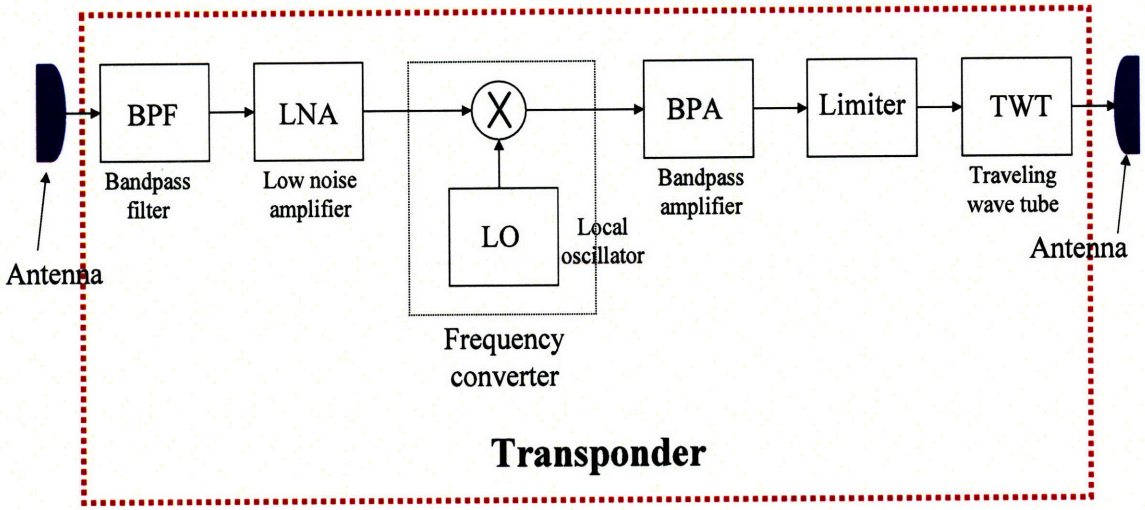

Figure 2-1: Block diagram of transparent repeater (consisting of one transponder) with receiving and transmitting antennas [7]

can be processed in a repeater at a time.

There are two types of repeaters: transparent repeaters and regenerative repeaters. A transparent repeater amplifies the signal and performs frequency conversion only. It is also called a bent-pipe repeater and Globalstar is an example. Fig. 2-1 shows the block diagram of a transparent repeater consisting of one transponder. The signal from the receiver antenna is fed into a bandpass filter (BPF) and a low noise amplifier (LNA). The BPF filters out unwanted noise outside the signal bandwidth and the LNA amplifies the weak received signal. The uplink frequency is converted to downlink frequency by mixing the amplified uplink signal with a local oscillator (LO). A bandpass amplifier (BPA) amplifies the signal in a small scale, followed by a high power amplifier (HPA) such as a traveling wave tube (TWT). Before the TWT, a limiter constraints the signal power below some level, to avoid the nonlinear distortion of the TWT (which is explained later).

The disadvantage of transparent repeaters is that noise and other distortion effects are amplified along with the uplink signal. Hence, a transmitted signal experiences two-way channel effects before reaching the intended receiver. Regenerative repeaters solve this problem by adding demodulation, baseband processing and re-modulation

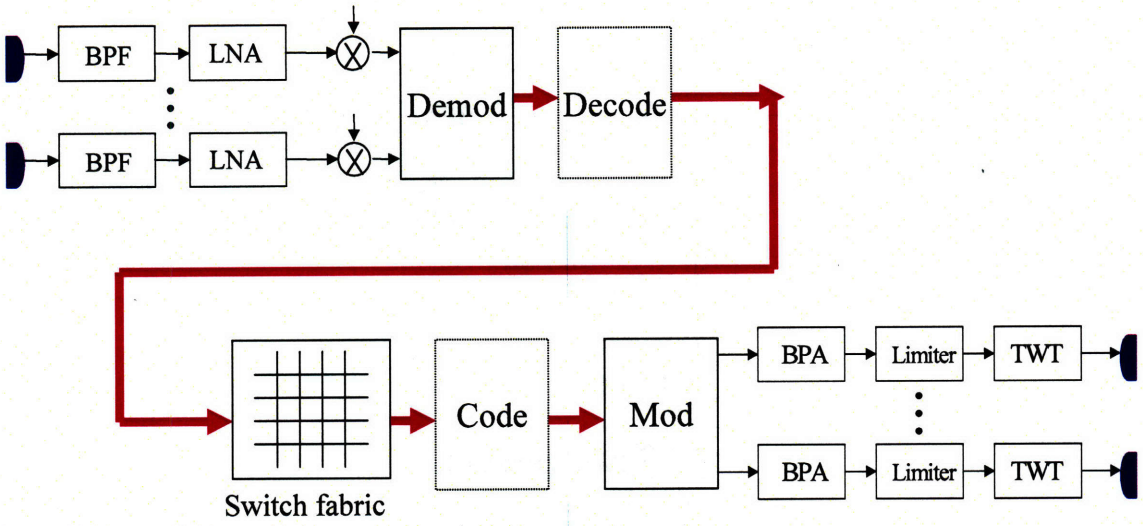

Figure 2-2: Block diagram of regenerative repeater (consisting of multiple transponders) performing on-board signal processing and switching [7]

(Fig. 2-2). A demodulator extracts the baseband signal from the radio-frequency transmission carrier and a decoder reconstructs information bits from the baseband signal. A regenerative repeater performs onboard processing of information bits, such as baseband switching to desired transmit spotbeams (if a destination is on the Earth over the downlink) or routing to intersatellite links (if a destination is a neighboring satellite). Information bits are re-coded and re-modulated into transmission carriers. A downlink transmission rate can be changed from the uplink rate by using different coding rates while this is infeasible in the transparent repeater.

2.2.2 Power Amplifier

High power amplifiers (HPA) are used before the output signal is fed to the transmit antenna. There are two types of HPAs widely used in satellite systems.

Traveling Wave Tube Amplifier (TWTA)

A TWTA is a vacuum tube that transfers energy from an electron beam to a radio frequency (RF) signal. In Fig. 2-3, the input-output power conversion can be ap-

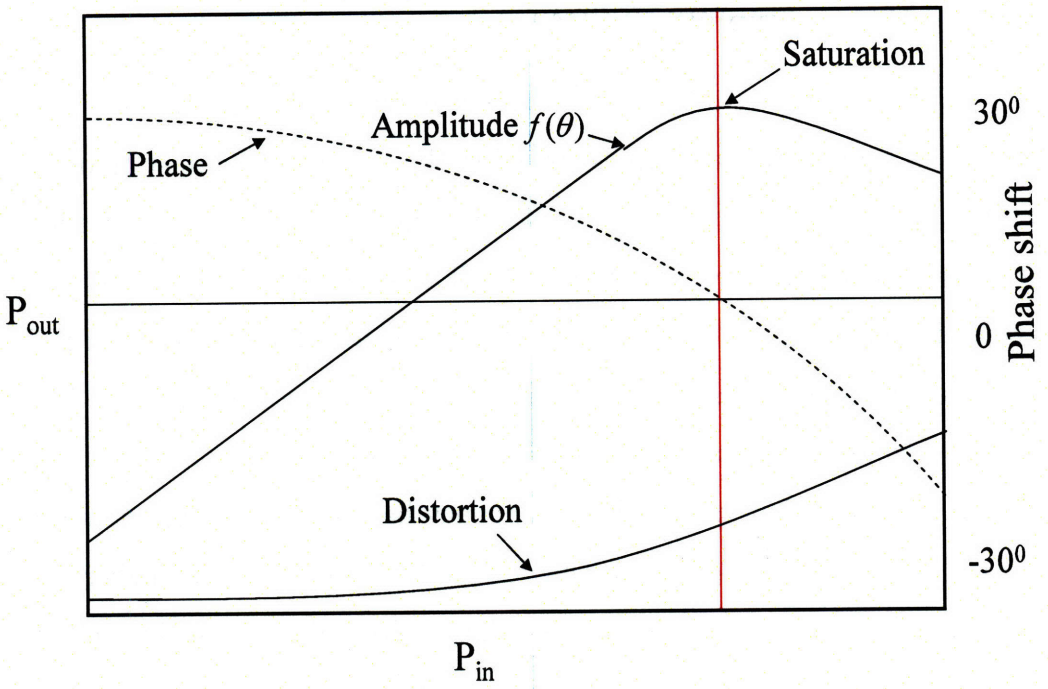

Figure 2-3: Input-output power conversion in TWTA [7]

proximated as linear under a threshold, called the saturation point. The saturation point gives the maximum input power that preserves the channel linearity. In a linear channel, the output of a combination of simultaneous inputs is equal to the sum of the output of the each individual input. If the input signal power ranges beyond the saturation point, the output signal does not show the same shape as the input and can cause amplitude and phase distortions, resulting in nonlinear channel characteristics. Channel nonlinearity makes signal detection and error recovery hard (e.g., channel equalization may be deployed to compensate signal distortion) and is better to be avoided. Thus, a TWTA is operated up to the saturation point, which gives the highest efficiency while ensuring channel linearity. In addition, when different frequency carriers are fed into a TWTA beyond saturation, the output has noise at the frequency that is a linear combination of input signal frequencies, which is called intermodulation noise. Thus, a TWTA is usually operated with a backoff in the range of $3 \sim 10$ dB under saturation. A typical TWTA with a helix can achieve output power less than 2.5 kW.

Solid State Power Amplifier (SSPA)

A SSPA utilizes microwave integrated circuits usually followed by the small radiating elements of an array antenna. Signals are fed to the elements by adjusting phases and amplitudes. Compared to a TWTA, a SSPA provides better linearity and reliability. A SSPA has a greater linear region than a TWTA, followed by sharp cut-off, because each radiating element has low output power. Due to a large number of radiating elements, failure of a few elements may be negligible. However, it is hard to scale up to the order of 100 W at low losses with the current technology. This is why TWTAs have been widely used in high frequency bands where a high power margin is required. It is reported that gallium nitride (GaN) [43] can achieve higher efficiency than other widely used materials such as silicon (Si), gallium arsenide (GaAs) and indium phosphide (InP). GaN is expected to be utilized for high-frequency (Ka band and beyond) high-power (more than 100 W) SSPAs with power efficiency up to 40 $\sim$ 60 % (compared to 20 % efficiency of TWTAs) [36]. By the nature of an array antenna with a large number of antenna elements, each of which is fed by a HPA, a SSPA is frequently used for an array antenna.

2.2.3 Modulation and Error Correction Codes

Currently, almost all satellite communications are digital since digital communications have low error rates through error detection and correction, and thus provide robustness to the channel noise that is a big problem in analogue communications. In all modern digital communications, coding and modulation techniques are used for efficient and reliable transmissions. Coding and modulation affect the relation between the signal-to-noise ratio (SNR) and bit error rate (BER), and influence the data transmission rate.

Digital modulation maps bits into symbols, shapes signal spectrum in the limited bandwidth, and translate the carrier frequency. A digital modulator produces analog waveform signals according to a sequence of discrete symbols from a finite number M of alphabets, into which binary bits from a source or coder are mapped. The

simplest scheme to implement for M-ary modulation in a satellite system is M-ary phase shift keying (M-PSK). All symbols have equal symbol energy and equal number of nearest neighbor signals. However, in satellite communications, where both power and bandwidth are precious resources, bandwidth efficient modulation (BEM) such as M-ary quadrature amplitude modulation (QAM) is considered for use. Nonlinearity of satellite channels imposes difficulty on the use of multi-layered envelop modulation schemes. Equalization is necessary to eliminate memory over the received symbols and intersymbol interference (ISI).[5]

Forward error correction (FEC) coding achieves a coding gain that reduces the amount of required transmission power for the desired BER, at the cost of broader bandwidth or lower data rates. Additional bits are assigned to information bits for error detection and/or correction. The ratio of the number of information bits to that of coded bits is called a code rate. The smaller a code rate is, the larger bandwidth is required. There are two types of codes widely used: block codes and convolutional codes. A systematic block encoder adds parity bits to a block of information bits to form a codeword. Parity bits are decided according to linear combinations of an incoming block of information bits. A convolutional code uses shift registers and adders, which generate coded bits from a segment (whose length is called a constraint length) of information bits. It is known that the optimum decoding method for convolutional codes is the Viterbi algorithm. A convolutional code with a Viterbi decoder is popular in current communication satellite systems due to its large coding gain and simple decoder structure that is feasible up to high rate transmission. For example, in the Globalstar system, the L band mobile uplink uses the convolutional code of rate 1/3 and constraint length 9, and the S band mobile downlink uses the convolutional code of rate 1/2 and constraint length 9. The Iridium system deploys the convolutional code of rate 3/4 and constraint length 7 for mobile user links, and rate 1/2 and constraint length 7 for intersatellite link transmission and feeder links between satellites and gateways.

[5]Due to channel degradation, a symbol of the received signal can spread over neighboring symbols, which degrades the decoding of the symbols. This is called intersymbol interference.

2.2.4 Frequency Reuse in Multibeam Satellites

A multibeam satellite can transmit many narrow spotbeams over multiple parallel transponders. The satellite coverage area is serviced by a number of small spotbeams, which can be considered as cells in a cellular system. By assigning the same frequency to non-adjacent spotbeams, one can reuse frequency and increase system capacity.

Since the signals in the same frequency can cause mutual interference (co-channel interference, CCI), techniques of mitigating interference are necessary. A conventional method in the cellular system is to separate co-frequency cells far enough, for interference to be negligible. Sometimes adjacent cells are grouped into a cell cluster, within which a different frequency is assigned to each cell. The frequency reuse factor, which represents how many times a frequency can be reused in the satellite coverage area, is given as the satellite coverage area divided by the cell size and the number of cells inside a cluster. Since the system capacity increases as the frequency reuse factor increases, the key point in the satellite cellular system is to decrease the cluster size and the cell size by the use of narrow spotbeams. An advanced satellite system (e.g., a military satellite) can suppress interference between close same-frequency signals by deploying a phased array antenna and locating nulls to other signals if necessary.

2.2.5 Multiple Access

In the satellite coverage area, many users compete for uplink access to the satellite transponder. A multiple access scheme divides the channel into small timeslots, frequency bands, or orthogonal codes, so that many users can share the channel without interfering with each other.

- Frequency division multiple access (FDMA): The frequency band is divided into small subbands, and each user sends a signal over a different frequency carrier. The receiver in the satellite filters the desired subband and reconstructs the signal. Since multiple signals are sent continuously at the same time, there can be a problem of adjacent channel interference (ACI) due to non-perfect channel

filters. To limit ACI, guard bands are inserted between frequency channels as much as 10 % of the channel bandwidth, and uplink power control is required to avoid the situation where strong power signals dominate other weak signals. In FDMA there can be intermodulation noise due to nonlinearity of satellite channels. When different frequency carriers are fed into nonlinear power amplifiers, the output has noise at the frequency that is a linear combination of input signal frequencies. In order to reduce the intermodulation noise, output power is backed off from full saturation, which results in the loss of power efficiency in the satellite transponder.

- Time division multiple access (TDMA): Transmission time is divided into small slots, which are assigned to different users. Each signal has a same carrier frequency over the full allocated bandwidth of the satellite transponder, and thus the satellite transponder processes one signal at one time. Since a user can send signals only at pre-assigned timeslots, clock synchronization is required and guard time intervals are inserted between timeslots. TDMA can support high rate digital transmission in a cost-effective way. An example of a satellite system using TDMA is Iridium.

- Code division multiple access (CDMA): All stations/users transmit at the same time over the same frequency band by using different signature sequences, called codes. By the use of codes, receivers can distinguish desired signals from others and the time-delayed version of the signal. To meet this requirement, a pseudo noise code is used, which gives approximate characteristics of random signals, so that orthogonality to other codes and its delayed version can be assured. Since the code is very fast in the time domain and occupies wide spectrum in the frequency domain compared to the signal, it spreads interference all over the band and thus reduces the interference level for the signal band. This is why CDMA is called spread spectrum transmission. An example of a satellite system using CDMA is Globalstar.

In all the multiple schemes of TDMA, FDMA and CDMA, every user is assigned to a fixed channel that is orthogonal to each other, so that interference is avoided. In the random access, users can transmit signals at any random time. Due to the possibility of collision between users, random access can be useful when there are a large number of users, sending short signals occasionally. One of its practical applications has been for VSAT (very small aperture terminals), which provides satellite communications between computers and small terminals.

The Aloha network was developed at the University of Hawaii in 1970 for communicating the central server with remote terminals via satellite and terrestrial wireless transmission. In the Aloha protocol, every user transmits randomly over the same frequency band either anytime (unslotted Aloha) or at the start of any time slot (slotted Aloha). If packets collide, they are backlogged and retransmitted randomly later. Since users in the Aloha do not need to wait for their pre-assigned channel as in the TDMA, the Aloha can achieve lower delay in the stabilized system, i.e., when the traffic arrival rate is less than the maximum throughput that the Aloha can maintain. However, the collisions between random accessed messages lead to a low throughput performance. The maximum throughput is only $1/e = 36.8$ % for the slotted Aloha or $1/2e = 18.4$ % for the unslotted Aloha, compared to 100 % of TDMA without any collision.

The poor throughput of the random access can be improved by resolving and reducing collisions at the cost of increased delay and system complexity. Under a simple splitting algorithm, the set of colliding users is split into subsets, and only one of the subsets can transmit in the next slot. Until the collision is resolved, the subset is split again and again. Anther variation is to monitor other users and to send messages only if others do not. This strategy is called carrier sense multiple access (CSMA). A reservation scheme can increase the throughput by sending short packets ahead of real data and reserving longer slots for the data. Reservations can be made by a random access of small packets or in a round-robin manner over small TDMA slots. With reservation, a higher throughput can be obtained due to less waste of idle

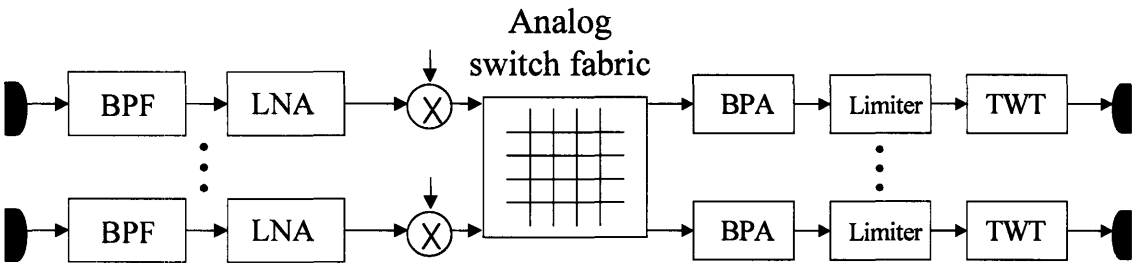

Figure 2-4: Analog switching without onboard signal processing [7]

or colliding time than the Aloha system.

2.2.6 Switching / Routing

Switching in a multibeam satellite interconnects the input signal from the uplink receive antenna to the transmit antenna in the downlink. Since interconnection is performed in cycle, input traffic is stored in a buffer and then transmitted in burst. Thus, satellite switching in practice is coupled with digital TDMA, called SS-TDMA (satellite switched time division multiple access). Switching can be performed by a programmable switch matrix that interconnects uplink and downlink beams, and the distribution control unit (DCU) that controls the sequence of connection states between inputs and outputs in time.

When the number of beams is small, an analog switching can hop the signal from a transponder to the other without on-board signal processing. Fig. 2-4 and 2-2 compare analog switching without onboard processing and digital switching with onboard processing.

Since LEO and MEO satellites have a small coverage area per satellite due to low altitudes and the constellation topology changes continuously, routing of traffic can be performed over inter-satellite links (ISL), e.g., in the Iridium. Each Iridium satellite uses up to four ISLs over Ka band at 23.18 - 23.38 GHz. Since there is no atmospheric attenuation of the signal through the ISL, the link power design only considers the free space loss, and it is highly desirable to use high frequency band and even optical

links for ISL. To establish transmissions over ISL, a satellite transmits a beacon signal, and thus, the receive satellite should maintain its receive antenna toward the transmit satellite precisely. The acquired beacon signal is used for the tracking of the satellite afterward. Future data satellites will have IP (Internet Protocol, explained in the next subsection) routers for header reading and switching.

2.2.7 TCP/IP

TCP/IP (Transmission Control Protocol and Internet Protocol) is the most dominant packet-oriented protocol in the Transport and Network Layers of the Internet. IP routes data packets over multiple networks from sources to destinations, using the IP header that is attached to the data packet and contains the information necessary for routing, such as type of service, total length, source IP address, and destination IP address. IP is an unreliable and best-effort protocol. Packets can be lost on the way and arrive out of sequence or in duplicate. Data are reconstructed from out-of-ordered packets in the destination buffer. Lost packets can be recovered by a higher layer protocol such as TCP that initiates retransmission from the source.

TCP is implemented on top of IP and achieves reliable end-to-end transmission by error control and retransmission of lost packets. The receiver sorts the received packets in the correct order by checking the sequence numbers that TCP assigns to the packets. TCP also provides a 16-bit check sum in the TCP segment header, in order to detect transmission errors. Acknowledgements (ACK) are sent back to the source from the receiver if the data delivery is completed and error-free. There are two cases when the source does not receive an ACK: when errors are detected in the received data or when data never arrive at all. The source retransmits the packet if it receives no ACK within some pre-set time interval, called a timeout.

Another important function of TCP is congestion control, to limit the amount of traffic entering the network below network capacity. Congestion control is performed by changing the transmission window size, which is the maximum amount of traffic the source can send in a single roundtrip ACK time. The system throughput is

then given as "window size/delay." TCP can increase the window size when the acknowledgements of the received packets from the destination are received. If no acknowledgement is received within a timeout period, TCP interprets this as traffic congestion in the network and reduces the window size (thus the system throughput). After sending packets and before receiving acknowledgements, TCP would wait for a round trip delay and an ACK to increase the window size.

TCP has been designed for low bit error rates and short delays of terrestrial fiber links. The long propagation delay over satellite links makes it hard to run TCP properly because it slows down the transmission of acknowledgements and thus wastes time and bandwidth. In addition, TCP cannot distinguish packet drops due to transmission errors over time-varying satellite channels with those due to network congestion. This degrades the TCP performance over satellite links by unnecessarily decreasing the window size even without congestion. There are efforts to solve this problem, including modification of the existing TCP, such as a large window size, or a design of new protocols.

One may split the TCP connection into terrestrial and satellite links, and run a proxy over satellite links. The proxy is designed to maximize the efficiency of the Transport Layer control over satellite links, for example, by distinguishing packet losses due to congestion and transmission errors. A router near the satellite link sends acknowledgements for the data and "spoofs" the source as if the transmission delay is short. The router then should read the acknowledgements from the destination node and retransmit any lost packets. This is called TCP spoofing. TCP spoofing is a practical solution for running TCP over satellites in current systems. However, it needs a lot of resources at the router near the satellite link to terminate all TCP connections and reinitiate a custom connection across the satellite link. And it is vulnerable to unexpected network changes or failures. Gilat Satellite Networks that use VSAT (very small aperture terminal) satellites for broadband communication services deploy "TCP acceleration," a type of TCP spoofing, with which most acknowledgements are handled locally by software at the hub and VSATs [24].

2.2.8 Topology Change and Handover in Networks of Satellites

When satellites (LEO or MEO) move, a channel in service can be out of connection, and then handovers, satellite-to-satellite and/or beam-to-beam, occur. Different from terrestrial wireless networks, user movement can be ignored in the satellite networks because of the large coverage and high velocity of satellites.

Handover decision is made based on channel conditions and the constellation topology. In a forward handover procedure, a handover request is submitted directly to a new satellite by a mobile user. The user maintains the old channel until a new channel is acquired. The handover break, which is the time interval between the initiation and the completion of the handover procedure, is shorter than a backward handover, where a handover request is sent through the old satellite. Some users may see multiple satellites sometimes and the CDMA Globalstar system with Rake receivers takes advantage of this chance. A Rake receiver detects the time-delayed versions of the signal in multipath environments and improves the system performance by combining all signal energies. When a user receives multiple line-of-sight signals from multiple satellites, a Rake receiver can combine all the signals and exploit satellite diversity as well as seamless satellite handover.

2.3 Remarks on Technologies

Thus far, we have reviewed some examples of commercial communication satellites and the main technologies for communication payloads. Here, we discuss what technologies are assumed to use and how they are modeled in the thesis.

- We consider a multibeam satellite with onboard processing of multiple signals. We assume the use of regenerative transponders, and focus on downlink transmission assuming that signals are re-modulated and switched onboard. The number of onboard modulators imposes a critical constraint to the number of

simultaneous active beams, and the two numbers are assumed to be identical throughout the thesis. Since every beam is assumed to carry a single signal different from other beams, the number of active beams is also equal to that of signals delivered at each timeslot. If transparent repeaters with analog switching are considered in the study, the number of TWTA decides that of active beams and signals.

- As described in the outline of Chapter 1, the thesis analyzes and compares the performances of the multiple beam antenna and the phased array antenna. The multiple beam antenna is equipped with TWTAs while the phased array antenna has SSPAs. The different implementation and physical characteristics of the two high power amplifiers result in different power constraints and eventually different performance results. Though the current technology of the phased array antenna with SSPAs is not sufficient to support a large antenna size in high frequency bands, it is expected that innovations of electronic and electro-optical phase-shifting/switching technologies can make the SSPA a better choice than the TWTA in the future, considering advantages of the SSPA: better linearity, better flexibility for beam shape/size and faster scheduling/cycling.

- In this thesis, we use Shannon capacity as a metric to allocate power. This assumes the use of an adaptive modulation scheme [8, 9], which can change the modulation size and thus the transmission rate according to traffic demand and channel conditions. In addition, advanced forward error correcting codes are assumed to be used (such as Turbo codes and low density parity check codes), to achieve the near-capacity data rate.

- In Chapters 3 and 4, we assume a cellular satellite system. Each cell is illuminated by a spotbeam and users inside a cell share the spotbeam by using a multiple access scheme (TDMA, FDMA or CDMA). On the other hand, in Chapter 5, we study the use of phased array antenna, which can synthesize more flexible antenna patterns than the multiple beam antenna, and thus, relax the

concept of the cellular system. Instead, we focus on individual user locations and combine the time-sharing scheme with interference-suppressed transmission for close-in users.

- Throughout the thesis, we consider the scenario where traffic demand exceeds system capacity. Admission control is required to insure system stability. In Chapter 3, we suppress the issues of delay and Transport Layer control and only focus on the Physical Layer resource allocation. In Chapters 4 and 5, we model admission control as a simple back-off parameter that throttles incoming traffic. This can be interpreted as a rate-based congestion control scheme that tells the traffic source how much data rate the satellite can accept.

- We mainly consider a single satellite system for a resource allocation and scheduling problem for the traffic that should be delivered to specific user destinations.

Chapter 3

Power and Beam Allocation Based on Traffic Demand and Channel Conditions

Narrow spotbeams in advanced satellite systems can project a high power density and thus can support high data rates for broadband data communications. In this chapter, we motivate the use of parallel narrow multibeams, and solve the optimization problem of multibeam power allocation based on traffic demand and channel conditions over satellite downlinks with power constraints. We maximize system performance and achieve reasonable fairness amongst users. We discuss the practical impact of a limited number of active beams and the need for multibeam scheduling.

In Section 3.1, we model the multibeam capacity over satellite downlinks. In Section 3.2, we suggest an optimum power allocation method and compare different cost functions of capacity and demand by trading-off between maximum total capacity and reasonable fairness. In Section 3.3, we analyze the power gain of optimum power allocation. In the following sections, we obtain modified results when we add the average delay constraint in the steady state (Section 3.4), and when the number of shared active downlink beams is smaller than that of service cells (Section 3.5). In Section 3.6, we summarize the chapter.

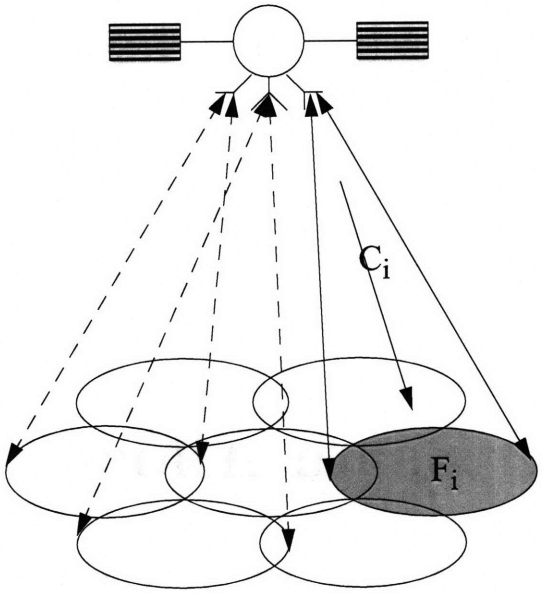

Figure 3-1: A multiple spotbeam satellite that provides capacity C_i for the i^{th} cell of demand F_i

3.1 Modeling of Downlink Multibeam Capacity

Diffraction theory [26] gives the relationship between transmitted and received power for satellite-to-earth links. Let P_i^t denote the transmit power and C_i the channel capacity in the i^{th} beam of a multibeam satellite to serve traffic demand F_i (which is the total amount including both new arrival and backlogged traffic) in the i^{th} coverage cell (Fig. 3-1). On-board transmission power is divided and shared among transponders to allocate capacities to cells. Multiple signals are transmitted simultaneously by using a multiple beam antenna or a phased array antenna. We assume for now that every beam is equipped with a transponder and carries a signal only for that beam. For a transmit antenna of diameter D, wavelength λ, and altitude of the satellite L, the mainlobe beamwidth illuminated by a diffraction-limited beam is $\frac{\lambda L}{D}$ and the received power P_i^r for a receiver antenna of diameter δ is given as

$$P_i^r = \left(\frac{\pi}{4}\right)^2 \frac{D^2\delta^2}{\lambda^2 L^2} P_i^t. \tag{3.1}$$

Within the i^{th} cell, using either superposition codes (which superimpose the signals

generated by different codebooks and send altogether after summing them) or a time-sharing scheme for Gaussian broadcast channels [13], we can achieve the band-limited Shannon capacity [14] of

$$C_i = W \log_2 \left(1 + \frac{P_i^r}{W N_0}\right) \quad \text{bits/sec}, \tag{3.2}$$

where N_0 is the noise power density and W is the bandwidth used. This capacity is achieved regardless of the number of uncoordinated receivers in the cell if we assume that all the receivers use the same size antenna.[1] There can be interbeam interference from the sidelobes of adjacent beams, to degenerate the Shannon capacity. However, we assume that interbeam interference is negligible because we consider very narrow beams over a large number of cells. In practice, we do not allow adjacent active beams and can locate active downlink beams far enough at each timeslot. Even in the case of adjacent active beams, the use of different bandwidth and/or polarization can suppress interbeam interference (at the expense of losing some efficiency). Spatially close co-channel interference will be considered in Chapter 5.

Since the capacity function of power is concave, its derivative with respect to power is monotonically decreasing (Fig. 3-2). To take advantage of this, we must not provide full power just for a single beam, but divide power and use multiple spotbeams with a small amount of power for each. When we have uniform allocation of power per beam (which gives the maximum capacity for a given number of multibeams by the nature of the concavity of the logarithm function) of $P_i^t = P_{total}^t/K$ where P_{total}^t is the total transmit power and K is the number of beams (which is equal to that of cells N for now), the maximum bandwidth efficiency on a satellite-to-earth path is

$$C/W = K \log \left(1 + \frac{\pi^2 D^2 \delta^2}{16 \lambda^2 L^2} \frac{P_{total}^t/K}{N_0 W}\right) \text{ bits/sec/Hz} \tag{3.3}$$

[1] Of course, when such schemes as joint decoding or maximum ratio combining amongst different receivers are deployed, we can achieve a higher capacity. However, this is unrealistic for the satellite downlink transmission scenario, where most receivers are indifferent to each other and it is impractical to construct any kind of network for joint processing.

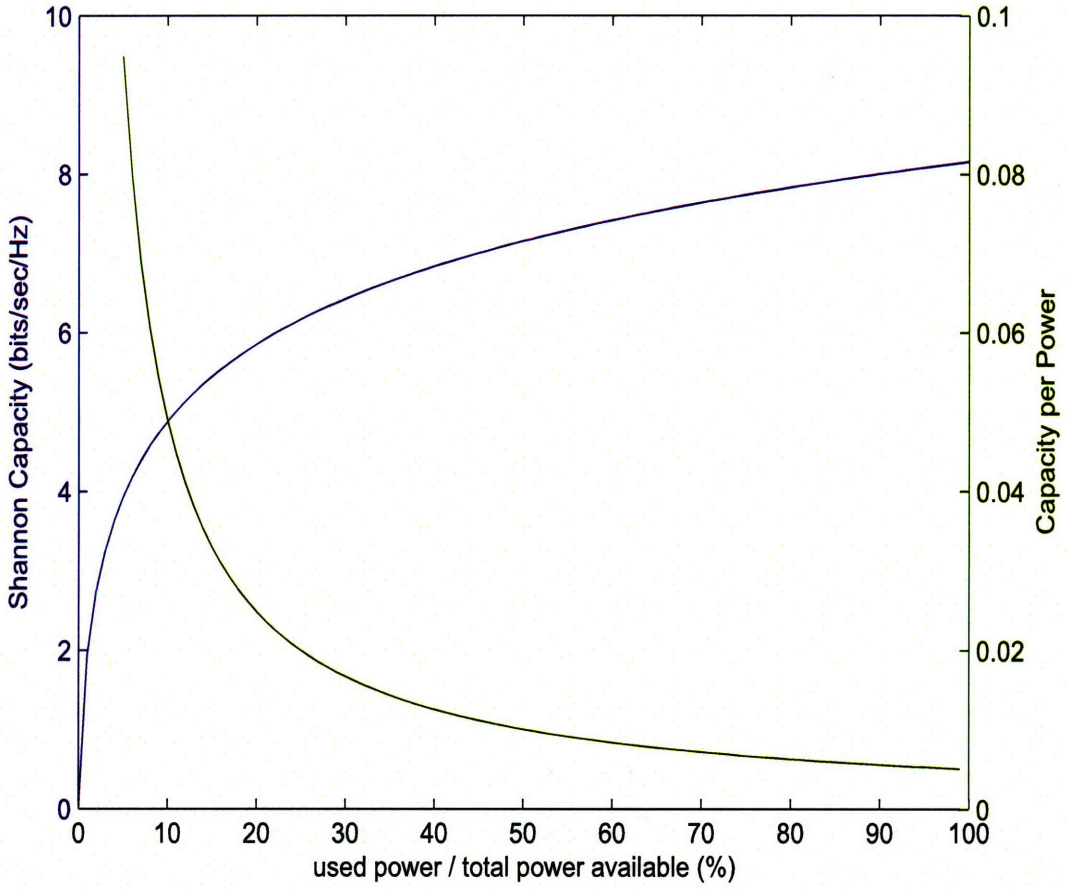

Figure 3-2: Concave capacity function (logarithm in this case; blue label in the left axis) of a single beam with respect to the used power out of the total power and its monotonically decreasing derivative (green label in the right axis)

$$\left(\sim \frac{\pi^2 D^2 \delta^2 P^t_{total}}{16 \lambda^2 L^2 N_0 W \ln 2} \text{ for large } K \right).$$

Eq. (3.3) indicates that the capacity gain of multibeams is monotonically increasing with the number of beams (Fig. 3-3). However, in practice, though the total capacity of a satellite is maximized, illuminating all areas all the time uniformly may cause a waste of resources because real traffic is nonuniform and time-varying. Thus, one needs to find optimum downlink power allocation across the parallel beams for any given user demand. Since the distribution of P^r_i is easily translated to that of P^t_i by

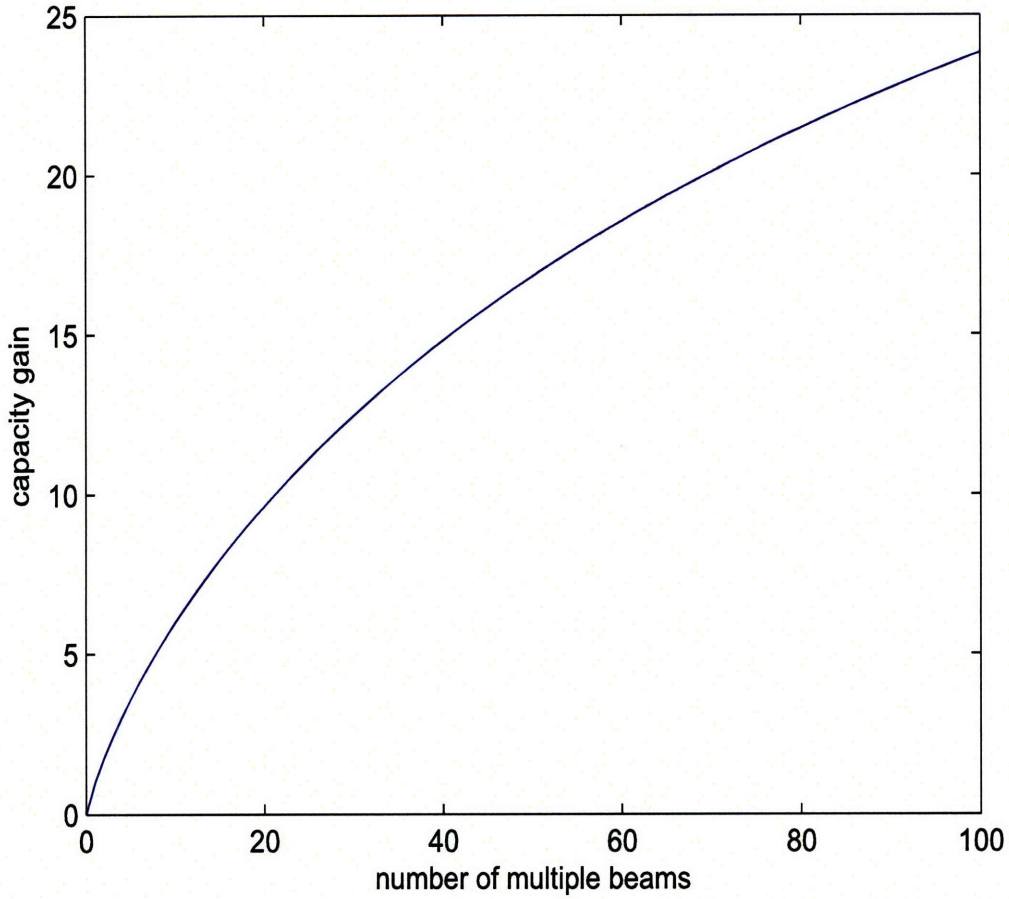

Figure 3-3: Capacity gain with uniform distribution of power (compared to a single beam) along the number of multiple beams (with parameters of Globalstar [25, 39])

the use of Eq. (3.1), we take into account only

$$P_i^r \equiv P_i \quad \text{and} \quad P_{total} \equiv \frac{\pi^2 D^2 \delta^2}{16\lambda^2 L^2} P_{total}^t \tag{3.4}$$

for this problem.

If there is cell-specific selective attenuation on the link, we can measure/estimate [9] and incorporate it in a modified expression of (3.1). The spotbeam that we consider in the future satellite is so narrow that the diameter of the cell can be as small as 50 $\sim$ 100 miles.[2] We can say that the correlation distance of fading events is larger than

[2]For an example of a current system, Iridium uses the spotbeams of diameters of about 500 miles.

this small spotbeam size in high frequency satellite channels, and assume uniform attenuation across each spotbeam. If we use the worst attenuation within a narrow spotbeam in our analysis, it yields a conservative bound for the performance, and in practice, the satellite would use short of individual measurement to each user in this way. When the i^{th} cell has signal power attenuation of α_i^2 (≤ 1) over the whole area of the cell, the received power becomes $\alpha_i^2 P_i$ and we can still infer P_i^t from P_i^r with the measured estimates of α_i^2.

3.2 Optimum Power Allocation

3.2.1 Performance Metrics

There are many metrics to evaluate system performance, such as maximum total capacity and fairness, and these different metrics may lead to very different system behaviors with different power and rate allocation. Thus, the choice of an appropriate metric is important for the study. Here, we want to match capacity C_i to accumulated traffic demand F_i as closely as possible, i.e., we want to minimize a general function of the difference between $\{C_i\}$ and $\{F_i\}$ across all cells.

Fig. 3-4 illustrates a simple two-user example, where the demands of two users are outside the capacity region (given by the total power and channel conditions). We compare three metrics of how to allocate power: maximum total capacity, proportional fairness and minimum square deviation. For maximum total capacity, one wants to maximize $C_1 + C_2$. The optimum allocation is given by the tangent point between the capacity region and the line of

$$C_1 + C_2 = C_{max} \tag{3.5}$$

over the plane of (C_1, C_2). Proportional fairness finds some constant a ($0 < a \leq 1$) that satisfies

$$C_i = aF_i \tag{3.6}$$

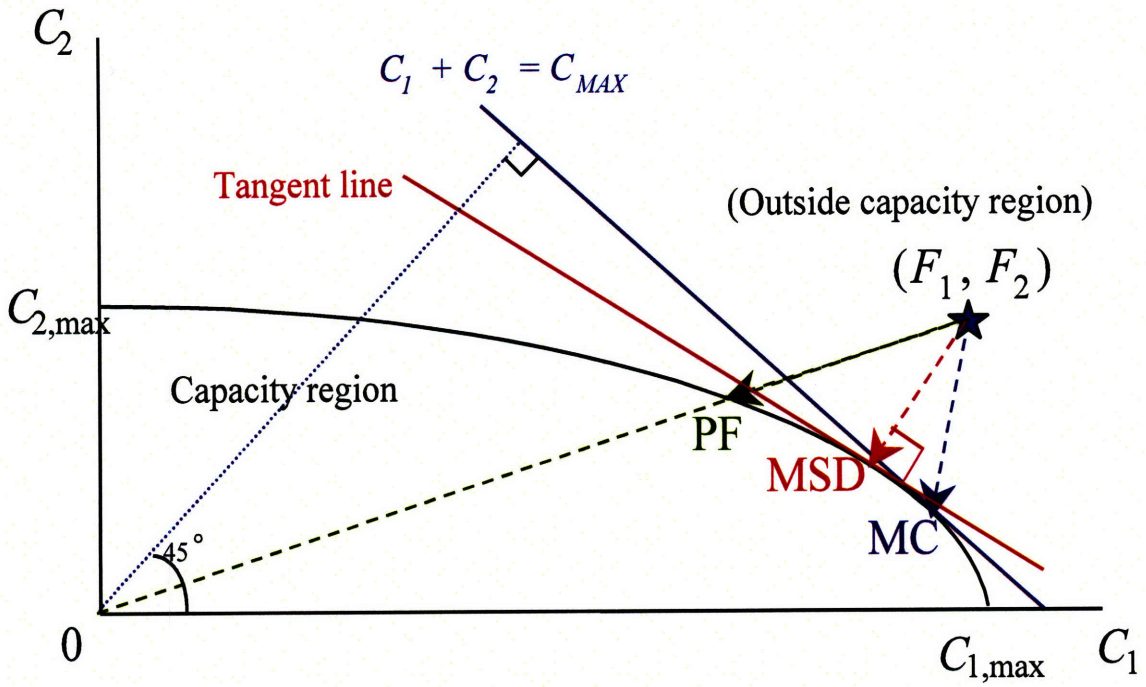

Figure 3-4: Comparison of three metrics of how to allocate power for the demand (F_1, F_2) outside the capacity region: maximum total capacity (MC), proportional fairness (PF) and minimum square deviation (MSD)

for $i = 1$ and 2. The optimum point is the intersection between the capacity region and the line of

$$\frac{C_1}{F_1} = \frac{C_2}{F_2}. \tag{3.7}$$

For minimum square deviation, i.e. minimizing $(F_1 - C_1)^2 + (F_2 - C_2)^2$, the optimum point gives the minimum Euclidean distance between the demand point (F_1, F_2) and the capacity region. As in this example, we will see that the metric of minimum square deviation provides a sensible compromise between total capacity maximization and fairness.

3.2.2 Square Deviation Cost Function

If we use a square deviation cost function, the problem can be modeled as

$$\text{minimize} \quad \sum_{i=1}^{N}(F_i - C_i)^2 \tag{3.8}$$

$$\text{subject to} \quad C_i = W \log\left(1 + \frac{\alpha_i^2 P_i}{W N_0}\right) \leq F_i \quad \text{for every } i, \tag{3.9}$$

$$\sum_{i=1}^{N} P_i \leq P_{total}, \tag{3.10}$$

$$\text{and} \quad P_i \leq P_0 \quad \text{for every } i. \tag{3.11}$$

In (3.9) we emphasize that we never use more power than required for traffic demand. The best case is when we have a trivial solution of $C_i = F_i$ for every i with $\sum P_i \leq P_{total}$. However, in many situations some or all beams may have capacities less than traffic demand with available power, which may lead to traffic congestion for the system. Solutions for this problem will involve accepting more delay, possible data routing on alternate paths and triggering Transport Layer congestion control mechanisms. Here we consider the case where the total amount of demand exceeds the total capacity, so that efficient allocation of power in the Physical Layer is mainly dealt with while the delay and Transport Layer control issues are suppressed. We consider the impact of selective signal attenuation at each coverage cell, mainly from rain that can cause slow (minutes or hours) but deep (about $10 \sim 20$ dB) fading over high frequency bands. As explained in Section 3.1, we assume uniform attenuation α_i^2 (≤ 1) across each narrow spotbeam. Condition (3.10) implies a constraint for the total power, and (3.11) implies that when each multiple beam antenna feed follows its own high power amplifier (HPA), such as a traveling wave tube amplifier (TWTA), every beam has a maximum transmit power constraint. On the other hand, if a phased array antenna is deployed, each beam is synthesized by adding array elements whose phases and amplitudes are adjustable, and we can provide as much power as we want in one cell (up to the total power of the array) by controlling the antenna-

patterning matrix and using solid state power amplifiers (SSPA). So with a phased array antenna, we can ignore the constraint (3.11).

The optimization problem is convex. Writing the Lagrangian function as

$$J(P_i) = \sum (F_i - C_i)^2 + \Lambda \left(\sum P_i - P_{total} \right) \tag{3.12}$$

in the case of phased array antenna (without constraint (3.11)) and differentiating with respect to P_i, we have

$$\frac{\partial J}{\partial P_i} = -2(F_i - C_i) \frac{\frac{W}{\ln 2} \frac{1}{WN_0}}{\frac{1}{\alpha_i^2} + \frac{P_i}{WN_0}} + \Lambda = 0. \tag{3.13}$$

Then, the optimum beam profile P_i satisfies

$$F_i - W \log \left(1 + \frac{\alpha_i^2 P_i}{WN_0} \right) = \frac{\Lambda N_0 \ln 2}{2} \left(\frac{1}{\alpha_i^2} + \frac{P_i}{WN_0} \right), \tag{3.14}$$

where Λ is a Lagrange multiplier and determined from the total power constraint. Nonnegative Λ means that Eq. (3.14) satisfies the restriction (3.9) of $C_i \leq F_i$. If we find $P_i > P_0$ in the case of multiple beam antenna, we will set $P_i = P_0$ by condition (3.11). The optimality of P_i is still valid and proved in the chapter appendix (Section 3.7) by using Theorem 4.4.1 in Gallager's textbook [22]. In general, Eq. (3.14) does not yield closed form solutions, and can be solved numerically to get P_i in terms of F_i. However, meaningful intuition can be drawn from closed form solutions by dividing cases at high and low signal-to-noise ratios. For example, at the low SNR region of

$$\frac{\alpha_i^2 P_i}{WN_0} \ll 1 \quad (= 0 \text{ dB}), \tag{3.15}$$

using

$$\ln(1 + x) \simeq x \tag{3.16}$$

for small x, we have

$$F_i - W \cdot \frac{1}{\ln 2} \cdot \frac{\alpha_i^2 P_i}{WN_0} = \frac{\Lambda N_0 \ln 2}{2\alpha_i^2}. \tag{3.17}$$

Since power P_i cannot be negative, the first order approximation is given as

$$P_i = \begin{cases} \frac{N_0 \ln 2}{\alpha_i^2}\left(F_i - \frac{\Lambda N_0 \ln 2}{2\alpha_i^2}\right) & \text{if } F_i > \frac{\Lambda N_0 \ln 2}{2\alpha_i^2} \\ 0 & \text{if } F_i \leq \frac{\Lambda N_0 \ln 2}{2\alpha_i^2}. \end{cases} \tag{3.18}$$

The optimality of P_i still holds even after negative solutions are discarded, which is proved in the same way as when $P_i > P_0$ and P_i is set to be P_0 (see Section 3.7).

For a high SNR of

$$\frac{\alpha_i^2 P_i}{W N_0} \gg 1 \quad (= 0 \text{ dB}), \tag{3.19}$$

(i.e., around or more than 10 dB), we have

$$F_i = W \log\left(1 + \frac{\alpha_i^2 P_i}{W N_0}\right) + \frac{\Lambda N_0 \ln 2}{2}\left(\frac{P_i}{W N_0}\right), \tag{3.20}$$

which is a monotonically increasing function of P_i. Thus, for given F_i we have unique P_i of the order of $O(F_i) < P_i < O(2^{F_i/W})$. If we use a truncated part of the Taylor expansion of

$$\ln(1+x) \simeq x - x^2/2, \tag{3.21}$$

we have the second order approximation of

$$P_i = \frac{N_0}{\alpha_i^2}\left[W + \frac{\Lambda N_0 (\ln 2)^2}{2\alpha_i^2} - \sqrt{\left(W + \frac{\Lambda N_0 (\ln 2)^2}{2\alpha_i^2}\right)^2 - 2F_i W \ln 2}\right]. \tag{3.22}$$

Fig. 3-5 compares these two approximations of Eq. (3.18) and (3.22) with the numerical solution of Eq. (3.14). This solution of P_i for given F_i is generic, i.e., applicable to any distribution of demand.

In the static case of two channels at a fixed time with $F_1 = F_2$ and $\alpha_1^2 < \alpha_2^2$, Fig. 3-6 shows that with smaller α_1^2 (heavier attenuation in a worse channel condition) the cell 1 attains smaller capacity because both functions of

$$f_1(P_i) = F_i - W \log\left(1 + \frac{\alpha_i^2 P_i}{W N_0}\right) \tag{3.23}$$

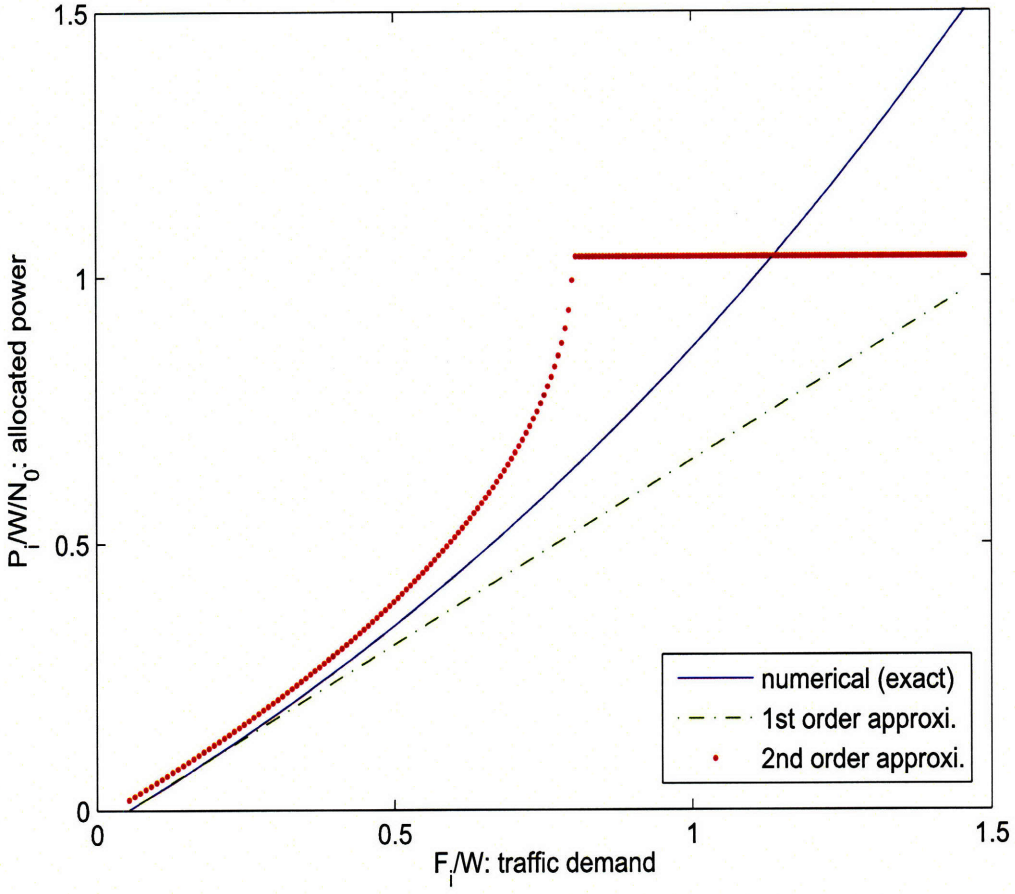

Figure 3-5: Optimum distribution of power P_i for demand F_i in Eq. (3.14) and its approximate closed-form answers (3.18) and (3.22)

and

$$f_2(P_i) = \frac{\Lambda N_0 \ln 2}{2}\left(\frac{1}{\alpha_i^2} + \frac{P_i}{WN_0}\right), \tag{3.24}$$

which determine the amount of P_i at the crossing point in Eq. (3.14), shift upward, to result in a larger deficit of $F_i - C_i$ for the same F_i. If all other parameters except channel conditions are identical, power is allocated such that the capacity of the worse conditioned channel is no larger than that of the better, which suggests to send more data through the better channel even if the amounts of traffic demand are equal in both channels.

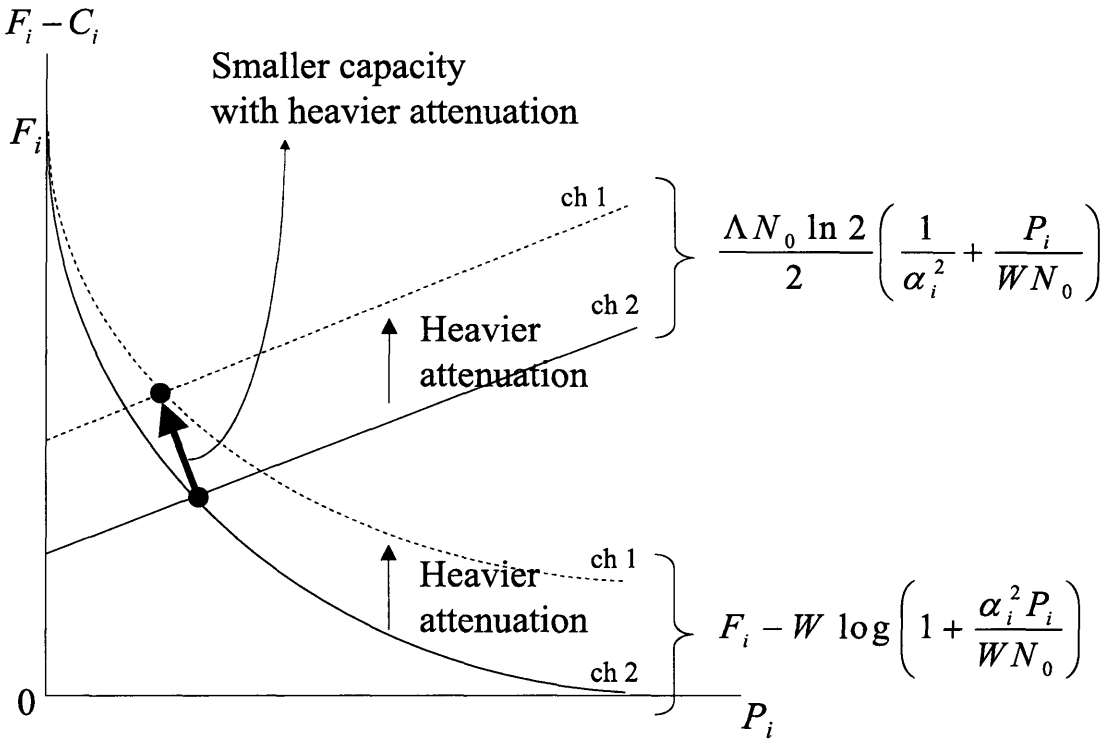

Figure 3-6: Illustration of difference of capacity with different signal attenuation in the static two-channel situation

3.2.3 Other Cost Functions

We now consider some other cost functions, such as first order, n^{th} order ($n \geq 3$) or linear scaling and compare different power distribution according to each cost function. We still assume that we provide no more power than required for traffic demand, i.e., $C_i \leq F_i$ for every i. With the first order cost function, we want to minimize

$$\sum_{i=1}^{N} |F_i - C_i| \tag{3.25}$$

for the same constraints as in the second order. For simplicity, let us denote the deviation as

$$\epsilon_i \equiv F_i - C_i. \tag{3.26}$$

Our problem is then given as

$$\text{minimize} \sum_{i=1}^{N} \epsilon_i \tag{3.27}$$

$$\text{s.t. } \epsilon_i \geq 0 \text{ (i.e., } -\epsilon_i \leq 0) \tag{3.28}$$

$$\text{and} \quad \sum_{i=1}^{N} \frac{2^{(F_i-\epsilon_i)/W}}{\alpha_i^2} \leq P'_{total}, \tag{3.29}$$

where the same power constraints as (3.9) and (3.10) of the second order are represented in terms of ϵ_i. With Lagrangian multipliers of Λ and μ_i, writing the Lagrangian function as

$$J(\epsilon, \Lambda, \mu) = \sum \epsilon_i + \Lambda \sum \left(\frac{2^{(F_i-\epsilon_i)/W}}{\alpha_i^2} - P'_{total} \right) - \sum \mu_i \epsilon_i \tag{3.30}$$

and differentiating with respect to ϵ_i, we have

$$\frac{\partial J}{\partial \epsilon_i} = 1 + \Lambda' \frac{2^{(F_i-\epsilon_i)/W}}{\alpha_i^2} - \mu_i = 0, \tag{3.31}$$

where $\Lambda' = -\frac{\ln 2}{W}\Lambda$.

By the Kuhn-Tucker condition [3], if $\epsilon_i = 0$ (i.e., if $C_i = F_i$), we have $\mu_i \geq 0$ and

$$1 + \Lambda' \frac{2^{F_i/W}}{\alpha_i^2} - \mu_i = 0. \tag{3.32}$$

If $\epsilon_i > 0$, we have $\mu_i = 0$, which leads to

$$1 + \Lambda' \frac{2^{(F_i-\epsilon_i)/W}}{\alpha_i^2} = 1 + \Lambda' \frac{2^{C_i/W}}{\alpha_i^2} = 0 \tag{3.33}$$

$$\Longrightarrow C_i = W \log_2 \left(\frac{\alpha_i^2 W}{\Lambda \ln 2} \right) \equiv \Gamma_i \quad \text{and} \tag{3.34}$$

$$\Lambda' = -\alpha_i^2 2^{-\Gamma_i/W}, \tag{3.35}$$

where Γ_i is a function of the channel condition. By combining (3.32) and (3.35) and removing Λ', we have

$$\mu_i = 1 - 2^{(F_i-\Gamma_i)/W} \geq 0, \tag{3.36}$$

which means that $F_i - \Gamma_i \leq 0$ when $\mu_i \geq 0$. Then, we obtain the solution of

$$C_i = \begin{cases} F_i & \text{if } F_i \leq \Gamma_i \\ \Gamma_i & \text{if } F_i > \Gamma_i, \end{cases} \tag{3.37}$$

where $\Gamma_i = W \log_2 \left(\frac{\alpha_i^2 W}{\Lambda \ln 2} \right)$ is determined by total power limitation and the channel condition of each cell. The first order cost function matches C_i to F_i perfectly until F_i meets the threshold Γ_i. The remaining power is distributed to the cell that requires more than Γ_i, so that the capacity is fixed at Γ_i regardless of traffic demand. Because we have

$$\sum |F_i - C_i| = \sum F_i - \sum C_i \tag{3.38}$$

with $C_i \leq F_i$, in this case we just want to maximize the total capacity. It is well known that the maximum total capacity over parallel Gaussian channels can be achieved by the water-filling solution [14]. While satisfying $C_i \leq F_i$, Eq. (3.37) represents water-filling since

$$C_i = \Gamma_i \tag{3.39}$$

is equivalent to

$$\frac{P_i}{W N_0} + \frac{1}{\alpha_i^2} = \text{ constant.} \tag{3.40}$$

If all cells have equal signal attenuation, Eq. (3.37) leads to a special case of uniform power distribution, as described in Section 3.1.

When we use a higher order cost function, say $n \geq 3$, of

$$\sum_{i=1}^{N} |F_i - C_i|^n \tag{3.41}$$

with all the identical restrictions, the result is a modified version of (3.14), which is written as

$$F_i - W \log \left(1 + \frac{\alpha_i^2 P_i}{W N_0} \right) = \left[\frac{\Lambda N_0 \ln 2}{n} \left(\frac{1}{\alpha_i^2} + \frac{P_i}{W N_0} \right) \right]^{\frac{1}{n-1}}. \tag{3.42}$$

We can expect that the difference of $F_i - C_i$ will be suppressed with more power in the high demand region and increased with less power in the low demand region, compared to the second order case. When we observe the threshold for nonzero power,

$$\left(\frac{\Lambda N_0 \ln 2}{n\alpha_i^2}\right)^{\frac{1}{n-1}}, \tag{3.43}$$

it is a monotonically increasing function of n with the assumption of the same value of Λ, which also assures that the high order function allocates more power to the higher demand cell while having the greater number of zero capacity cells for small demands.

Instead of using the deviation cost functions that depend on $F_i - C_i$, we can make C_i a scaled version of F_i, i.e.,

$$C_i = aF_i \tag{3.44}$$

in every cell for some constant $0 < a \leq 1$, so that ideal proportional fairness can be achieved since all cells are given the same proportion of capacities according to their demands. Power allocation and the scaling factor a are determined by numerically solving a set of nonlinear equations with the total power limitation, i.e.,

$$P_i = \frac{WN_0}{\alpha_i^2}(2^{\frac{aF_i}{W}} - 1) \quad \text{for every } i \tag{3.45}$$

and

$$\sum_{i=1}^{N} P_i \leq P_{total}, \tag{3.46}$$

which can be simplified according to the amount of demand, given as

$$P_i \simeq \begin{cases} \frac{aN_0(\ln 2)}{\alpha_i^2} F_i & \text{for low } F_i/W \\ \frac{WN_0}{\alpha_i^2} \cdot 2^{\frac{aF_i}{W}} & \text{for high } F_i/W. \end{cases} \tag{3.47}$$

We remark that the definition of "proportional fairness" in this thesis is slightly different from those used in the literature. According to the definition in Kelly *et al.*'s work [35], allocated capacity C_i is proportionally fair per unit demand if C_i is feasible

and if for any other feasible capacity C_i^*, the weighted aggregate of proportional changes is zero or negative, i.e.,

$$\sum_i F_i \cdot \frac{C_i^* - C_i}{C_i} \leq 0. \tag{3.48}$$

This allocation is known to maximize

$$\sum_i F_i \log C_i. \tag{3.49}$$

If we solve the Lagrangian function of

$$J(P_i) = \sum F_i \log C_i - \Lambda(\sum P_i - P_{total}), \tag{3.50}$$

the optimum power and corresponding capacity allocation is given by

$$\frac{F_i}{C_i} \cdot \frac{dC_i}{dP_i} = \Lambda, \tag{3.51}$$

which can be a form of

$$C_i = aF_i \tag{3.52}$$

only if

$$\frac{dC_i}{dP_i} = \text{ constant}, \tag{3.53}$$

i.e., a capacity is a linear function of power. However, in the general case of a nonlinear capacity with a finite bandwidth, it does not hold and this definition of "proportional fairness" gives a different result. Our proportional fairness can be obtained by solving a problem of

$$\max \min_i \quad \frac{C_i}{F_i} \tag{3.54}$$

and using the same argument as in Yang and Xu's work [56], where the authors solve a problem of

$$\max \min_i \quad SINR_i, \tag{3.55}$$

i.e., maximizing the smallest signal-to-interference-and-noise ratio, and show that it is optimal to equalize each user's downlink performance with

$$SINR_1 = SINR_2 = \cdots = SINR_N. \tag{3.56}$$

3.2.4 Comparison of Cost Functions

Fig. 3-7 shows the distribution of capacity for various cost functions along 20 cells that have a simple linearly distributed demand of

$$F_i = i \cdot \beta, \tag{3.57}$$

where β (> 0) is a slope of traffic distribution. The parameters of Globalstar are used as in Fig. 3-3. With $\alpha_i^2 = 1$ for every i, we first focus on the impact of different traffic distribution across cells. As we use a higher order cost function, more power is provided for higher demand cells while a lower order cost function gives relatively more power to lower demand cells. For example, in Fig. 3-7, the cubic cost function provides no power and no capacity for 7 lowest demand cells while the square cost function zeroes only one lowest cell and the first order cost function yields no zero-capacity cell. Since the square and cubic cost functions have power distribution patterns closer to that of linear scaling by serving higher demand cells better, they can be considered as proportionally fairer than the first order. However, higher capacity needs more power per bit due to concavity of the capacity function with a bandwidth constraint (logarithms in this analysis). This results in the lower total capacity across all cells when we use higher order functions. Total capacities per bandwidth of cubic and square cost functions are 55.93 and 71.19 bits/sec/Hz respectively in Fig. 3-7, while that of the first order is 78.50 bits/sec/Hz. Due to convexity of high-order cost functions ($n \geq 2$) with respect to allocated capacity, they should suppress the deficit in high F_i to minimize the total deficit across the cells, and thus allocate more power to more demanding cells to approach the behavior of proportional fairness.

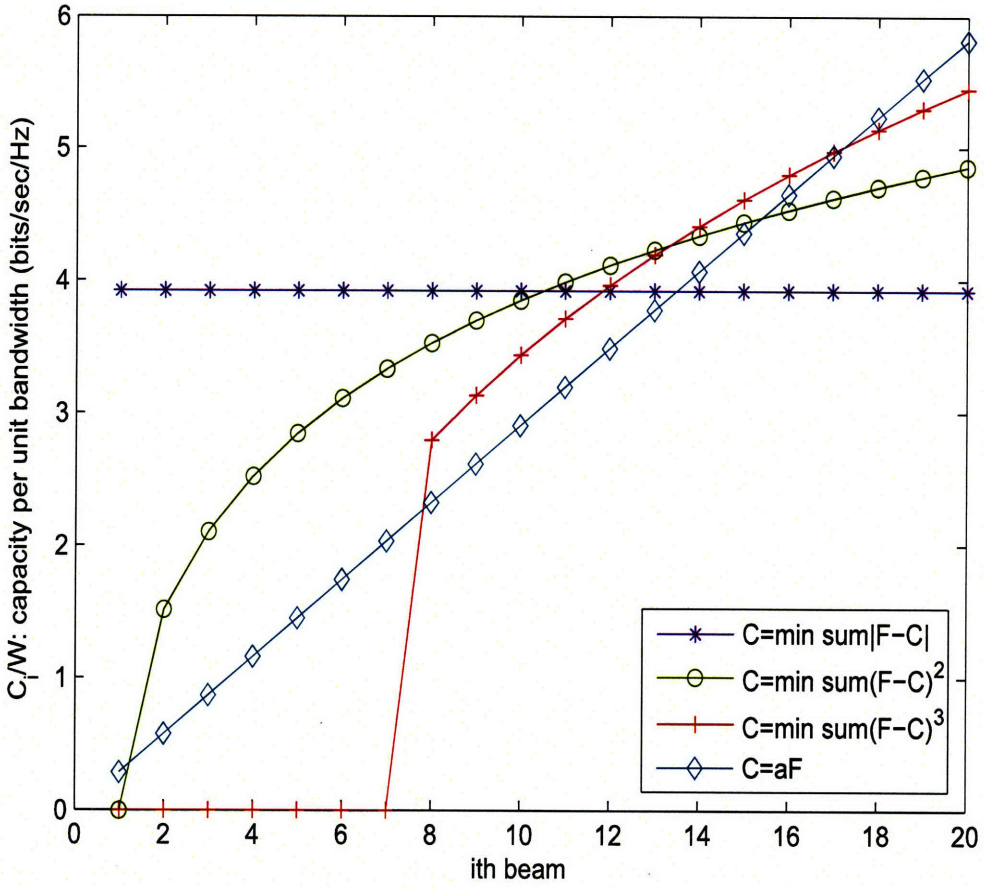

Figure 3-7: Comparison of cost functions in terms of capacity per unit bandwidth

The first-order function is linear and indifferent to the cell demand.

Therefore, we face the problem of trading-off between "maximum total capacity" and "proportional fairness." As performance metrics, maximum total capacity and proportional fairness suggest opposite directions of distributing power. If we want to maximize the total capacity, we should allocate a fixed amount of power regardless of actual traffic demand (if the demand is higher than some threshold for each cell), and lose proportional fairness. If the goal is to secure proportional fairness, on the other hand, we should allocate more power to channels with higher demand, but lose some total capacity because a higher capacity in a band-limited channel needs higher power per unit capacity. In our optimization problem, with more power for higher

demand in the case of the second or third order cost function, we can achieve more proportional fairness but lose some in total capacity, and the opposite trend holds in the case of the first order cost function. This result opens some possibilities of constructing a complicated cost function by combining lower and higher order cost functions as a sensible compromise.

Power allocation over different channel conditions also depends on the metric of choice. With maximum total capacity as a metric, we want to utilize better channels by following the water-filling strategy and allocating more power to them while ignoring worse channels. For proportional fairness, we put more power to worse channels to overcome channel degradation and to provide fairness among cells according to traffic demand. Fig. 3-8 compares allocated power for three metrics of minimum square deviation (2nd order cost function), maximum total capacity (1st order) and proportional fairness, with the same amount of traffic demand for every cell. In this example, power allocation by the second order cost function is very close to that by the metric of maximum capacity. However, the second order gives more power to worse channels and the threshold of attenuation[3] where the channel turns off is lower, compared to the first order. Thus, there is another trade-off between proportional fairness and maximum total capacity with different signal attenuation, and the second order cost function makes a compromise between proportional fairness and maximum total capacity.

In the dynamic environments where demand and channel conditions are time-varying, power allocation is updated every timeslot according to the fluctuation of demand and channel status, with a new Lagrangian multiplier Λ. Since Λ is dependent on the conditions of all cells, even if the demand and channel condition of one cell remain the same as in the previous timeslot, its power allocation may change as the circumstances in other cells change. Also, in practice, the power allocation method in presence of channel degradation must be coupled with link prediction schemes [9].

[3]The threshold of attenuation denotes the value of α_i^2 below which the channel condition is so poor that the channel loses any economic sense to utilize and no power is allocated, and thus the channel turns off.

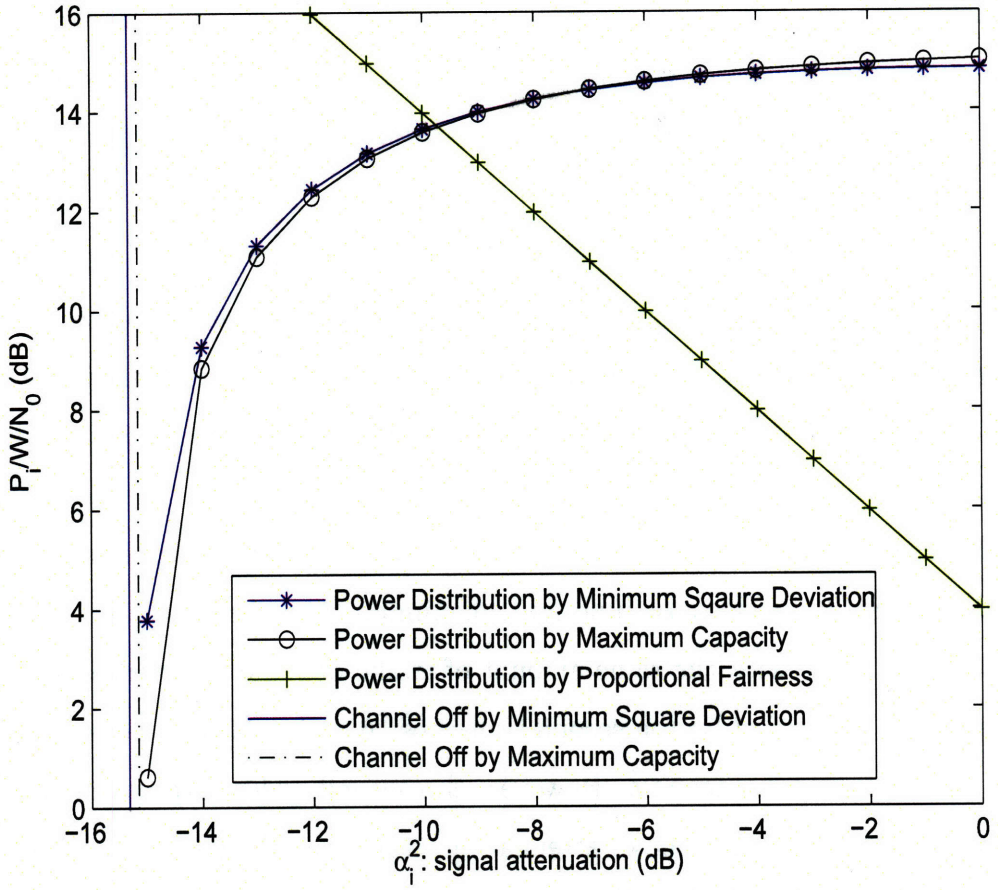

Figure 3-8: Comparison of power allocation according to signal attenuation, based on different metrics of minimum square deviation (2nd order cost function), maximum total capacity (1st order) and proportional fairness

This can be done by the transmitter alone based on estimates on a reciprocal channel from the receiver or from direct estimates by the receiver feedback in a return channel.

3.3 Power Gain by Optimum Power Allocation

By the nature of the capacity function in terms of power, we see that a satellite with parallel multibeams can make better use of a fixed amount of power to provide a higher capacity rather than with a time-sequentially scanning single beam only. When we deliberate on nonuniform and time-varying demand and the limited amount of on-

board power, it becomes important to allocate power optimally. Here, we compare the amount of total power spent for optimum power allocation with that for uniform allocation when both achieve the same square deviation cost $\sum_i (F_i - C_i)^2$. We define the power gain $g(N)$ of parallel multibeams with optimum power allocation over uniform allocation, as a function of the number of beams, N, given as

$$g(N) = \frac{NP_{uniform}}{\sum_{i=1}^{N} P_i} = \frac{NP_{uniform}}{P_{total}} \tag{3.58}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \{F_i - C(P_{uniform})\}^2 = \sum_{i=1}^{N} \{F_i - C(P_i)\}^2, \tag{3.59}$$

where $C(\cdot)$ is the band-limited Shannon capacity function, and $P_{uniform}$ denotes the required power if it is uniformly allocated to all cells to achieve the same cost as in the optimum allocation case. To take into account the waste of power used to serve low demand, we consider perfectly uniform power allocation, so that $F_i < C(P_{uniform})$ may happen.[4] Uniform allocation uses the total power of $NP_{uniform}$ and optimum allocation uses that of $\sum_i^N P_i = P_{total}$ (fixed in this case). In Fig. 3-9, for the same example as in Fig. 3-7, the power gain by optimum power allocation is more than 8 dB at $N = 100$. By allocating power optimally, we not only save total power by reducing the waste of power for small demand from the viewpoint of satellite operators, but also achieve reasonable proportional fairness from the viewpoint of users. The power gain also depends on the shape of traffic distribution, which is represented by a slope β of linear distribution in this case. For the same total amount of demand across the cells with 100 beams, Fig. 3-10 shows that the more unbalanced the distribution is, the more power gain can be realized because the optimum method can take advantage of nonuniformity of distribution by providing more (less) power to more (less) demanding cells.

[4]We note that there may be some cases where the perfectly uniform power allocation in (3.59) cannot be fulfilled. One example is when we have the small deviation cost of $\sum_i \{F_i - C(P_i)\}^2$ with extremely unbalanced F_i over the cells, which we do not consider here.

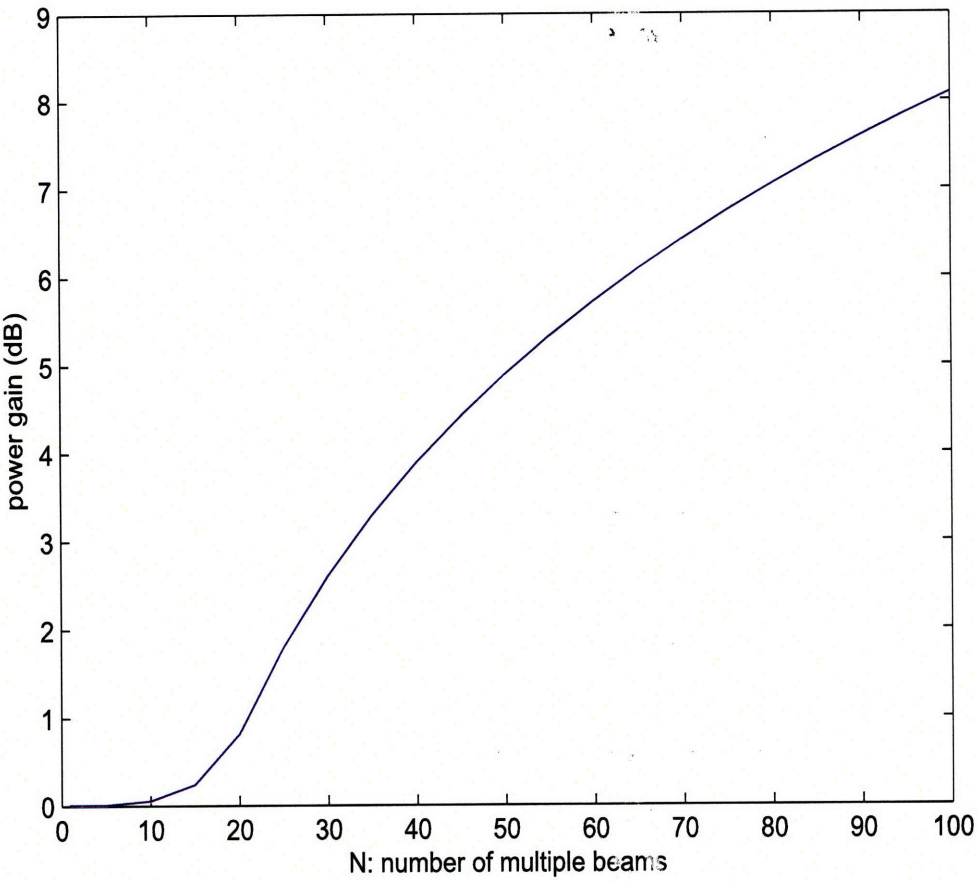

Figure 3-9: Power gain of parallel multibeams with optimum power allocation based on linearly distributed demand, compared to uniform power allocation, as a function of the number of multiple beams with fixed traffic distribution

3.4 Impact of Average Delay Constraints in Steady-State

In this section we add another important constraint, *delay*, to the optimum resource allocation problem. In practice, for many real-time applications such as video or audio conferencing, delay performance is as critical as error recovery. It is likely that in most cases, a multibeam satellite deals with heterogeneous real-time and non-real-time traffic. By extending the formulation on accumulated traffic demand F_i and

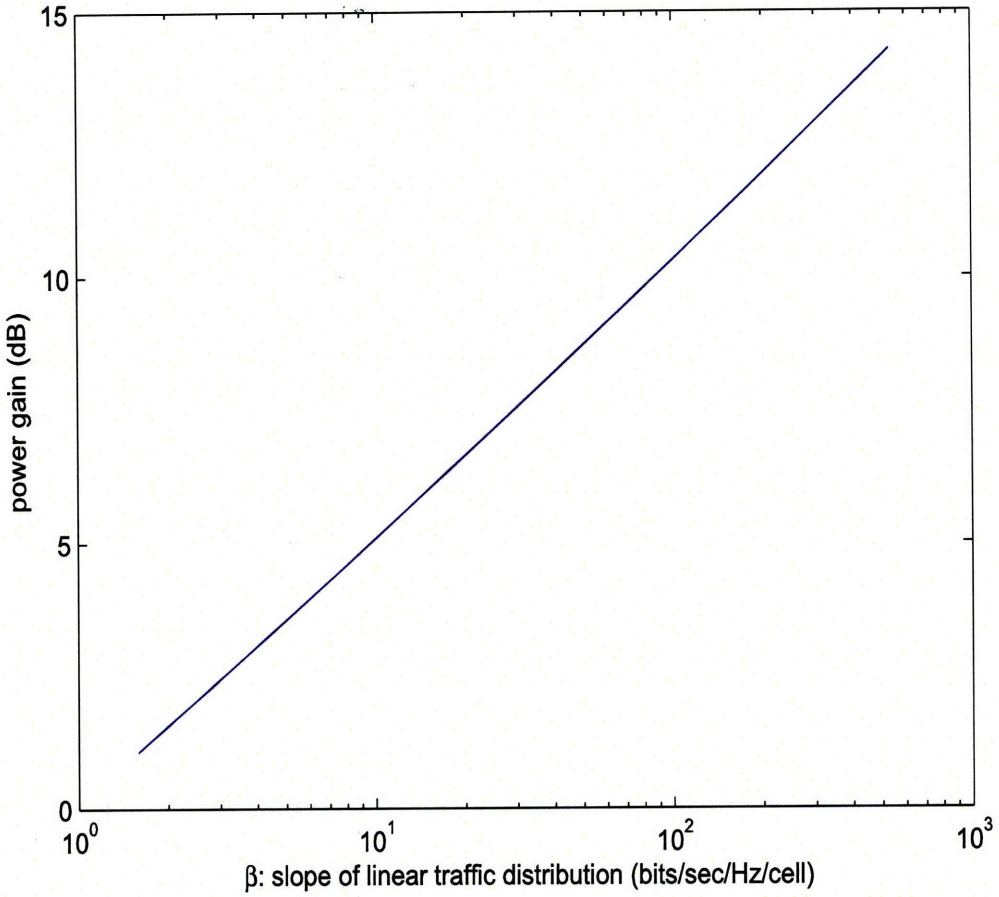

Figure 3-10: Power gain of parallel multibeams with optimum power allocation based on linearly distributed demand, compared to uniform power allocation, as a function of the slope of linear traffic distribution with $N = 100$ beams

capacity C_i in Section 3.2, we let A_i denote the average rate of incoming traffic and R_i the average rate of traffic removed from the queue and served (i.e., transmitted), respectively, in the steady state. We assume that every queue is stable, that is, $F_i < \infty$ with $A_i \leq R_i$.

We analyze the impact of a finite average delay on the transmit amount R_i. When there is no transmission error, we have $0 < R_i \leq C_i$ and the probability that the transmission is successful is $1 - e_i$, where e_i is the packet error rate (PER) over the link of the i^{th} beam. When there is a transmission error, we have $R_i = 0$ and the

transmission-failed demands are backlogged. The probability for this case is e_i. Thus, we have $R_i \leq (1-e_i)C_i$, considering both cases. Then, by Little's theorem [4], the steady-state time average delay in the i^{th} queue is given as

$$\bar{d}_i = \frac{F_i}{A_i} \geq \frac{F_i}{R_i} \geq \frac{F_i}{(1-e_i)C_i}. \tag{3.60}$$

We note that the above relation (3.60) holds for the steady-state averages of all the quantities. In particular, temporal traffic variation and channel conditions can be assumed to be quasi-static over the period of interest because the packet processing time and transmission deadlines are much shorter than the coherence time of signal attenuation due to rain, which is of the order of minutes or hours. And in general, accumulated traffic changes more slowly compared to channel conditions.

Suppose that the i^{th} beam has the average delay constraint of $\bar{d}_i \leq \Delta_i$ for $i = 1, ..., N$, where Δ_i (> 0) is a given average delay deadline and may be different from beam to beam. We focus only on the long-term average delay of each beam by assuming a "genie-aided" Transport Layer Protocol that properly serves the congested and backlogged demand as well as new incoming traffic. Also, in (3.60), we assume $A_i \simeq R_i$ in the steady state and $R_i \simeq (1-e_i)C_i$ by the use of error correction codes, thus the gaps are reduced in inequalities. Hence, we have an average delay constraint of

$$\frac{F_i}{(1-e_i)C_i} \leq \Delta_i \quad \text{or} \quad \frac{F_i}{(1-e_i)\Delta_i} - C_i \leq 0 \tag{3.61}$$

in terms of F_i and C_i. This constraint implies that the i^{th} beam has to secure at least some fraction $(\frac{1}{(1-e_i)\Delta_i})$ of demand for the capacity to meet the delay constraint. The fraction is determined by the delay deadline and PER over the channel. Intuitively, a shorter deadline with higher priority transmission leads to a larger fraction of capacity, and a larger PER in a worse channel condition also requires a larger fraction in order to overcome a poor link quality. A larger capacity requirement for the worse channel condition can lead to the loss of the total capacity because the water-filling principle shows that it is better to take advantage of a good channel condition by

allocating more power in order to maximize the total capacity of parallel Gaussian channels. Thus, we have a problem of trading-off between delay and total capacity over attenuated channels. We remark that the delay considered here is closer to the transmission delay F_i/R_i than to the queueing delay F_i/A_i. The result with the queueing delay is discussed in later chapters.

By adding the average delay constraint to the original minimization problem and applying Lagrangian multipliers $\{\kappa_i, i = 1, ..., N\}$, we have

$$F_i - C_i + \frac{1}{2}\kappa_i = \frac{\Lambda N_0 \ln 2}{2}\left(\frac{1}{\alpha_i^2} + \frac{P_i}{WN_0}\right) \quad \text{for } i = 1, ..., N, \tag{3.62}$$

from which we calculate power distribution as assuming every $\kappa_i = 0$, first. If we have

$$C_i \geq \frac{F_i}{(1 - e_i)\Delta_i}, \tag{3.63}$$

it satisfies the Kuhn-Tucker condition [3] with $\kappa_i = 0$. If we have

$$C_i < \frac{F_i}{(1 - e_i)\Delta_i}, \tag{3.64}$$

C_i and the corresponding P_i should increase with $\kappa_i > 0$. Thus, we set

$$C_i = \frac{F_i}{(1 - e_i)\Delta_i} \tag{3.65}$$

and recalculate, so as to meet the average delay constraint. Fig. 3-11 and 3-12 show an example of power allocation and the corresponding capacity increase, $\frac{1}{2}\kappa_i$ in Eq. (3.62), needed for the average delay constraint, as a function of deadline. For deadlines longer than some threshold (1.11 in this example), power is allocated in the same way as without the average delay constraint and no additional capacity is needed ($\frac{1}{2}\kappa_i = 0$). In some cases, average delay constraints may not be satisfied for lack of available power, which leads to blocking or dropping of the service. In this modeling and analysis, assuming quasi-static channel conditions and traffic variation, we have considered the steady-state case only, so that the average delay constraint is a simple

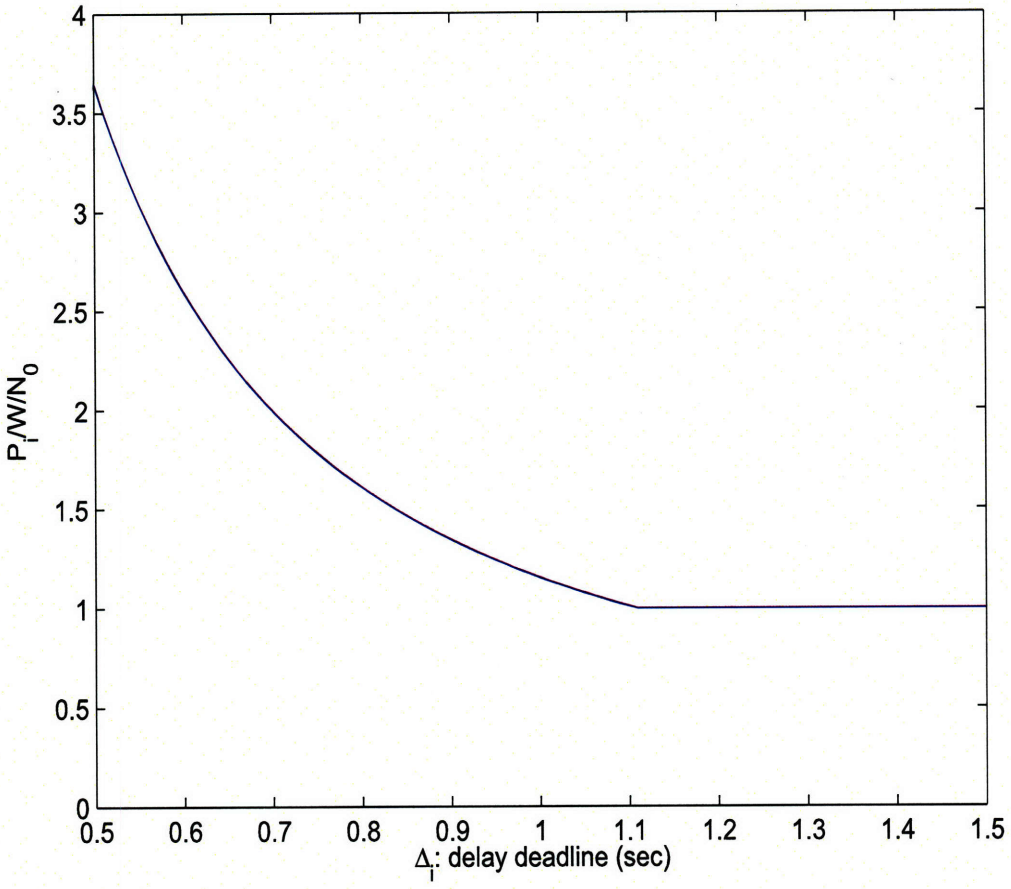

Figure 3-11: Power required for the delay constraint, as a function of delay deadline

form in terms of demand and capacity. This is to highlight the delay-constrained power allocation problem and provide insightful solutions. Time-varying fading and stochastic traffic loads are considered in Chapter 5. Nevertheless, this simple analysis with a queuing model indicates that for beams carrying traffic with shorter delay constraints, more power and capacity should be allocated. In our formulation, the beams with average delay constraints have priority to secure power and capacity resources over those without constraints. However, by differently weighing each term in (3.8), we can control the rank of queues, so that, for example, some important non-real-time services get more resources and are better served.

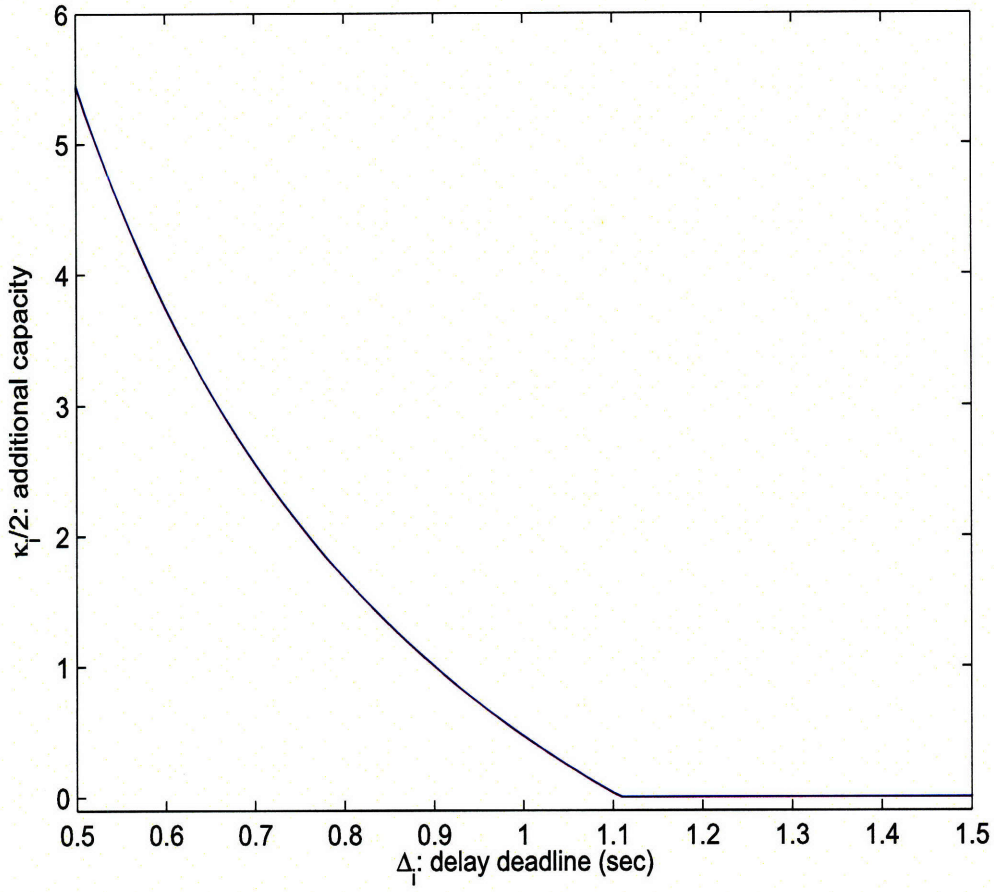

Figure 3-12: Additional capacity required for the average delay constraint, as a function of delay deadline

3.5 Impact of Shared Active Downlink Beams

Thus far, we have assumed that the number of cells in the satellite service area was the same as that of active downlink beams, i.e., a satellite could fully cover all its cells simultaneously. In the future, a satellite may have to use hundreds or thousands of cells to support small users with narrow spotbeams for high power density. It will be inefficient for a satellite to carry as many transponders, one for each spotbeam. So we need to find a way to efficiently share and schedule a smaller number of active downlink beams to cover a much larger number of cells in a coverage area.

Denote the number of available active beams as K and the number of cells as

N ($> K$). That is, at most K downlink beams can carry signals, generated by modulators on-board, to K cells and the remaining $N-K$ cells have zero capacities. The problem of finding which K cells should be serviced can be formulated as a binary programming problem with the same constraints of $C_i \leq F_i$ and total power limitation,[5] given as

$$\text{minimize} \sum_{i=1}^{N} |F_i - z_i \cdot C_i|^n \tag{3.66}$$

$$\text{subject to} \quad \sum_{i=1}^{N} z_i = K \text{ where } z_i \in \{0,1\}, \tag{3.67}$$

where z_i indicates whether the i^{th} cell will be covered or not, and we consider only $n > 1$. We may solve this problem by considering all $\binom{N}{K}$ cases and picking the optimum.

However, it is hard to solve this kind of integer programming problem and to get any insight behind the solution. Thus, we go back to the initial minimization problem by relaxing the binary condition of z_i, given as

$$\text{minimize} \sum_{i=1}^{N} (F_i - C_i)^n \tag{3.68}$$

$$\text{subject to} \quad 0 \leq C_i \leq \min\{C_i^0, F_i\} \text{ for every } i \tag{3.69}$$

$$\text{and} \quad \sum_{i=1}^{N} \frac{2^{C_i/W}}{\alpha_i^2} \leq P'_{total}, \tag{3.70}$$

where we focus on C_i instead of P_i. In (3.69) and (3.70), power limitation conditions are also represented using C_i and C_i^0, the maximum capacity of the i^{th} cell from the constraint (3.11). Using Lagrangian multipliers μ_i for $-C_i \leq 0$ and ν_i for $C_i \leq C_i^0$

[5]We suppress the delay constraint here, so as to focus on the solution based on traffic demand and channel conditions only.

respectively,[6] the corresponding Lagrangian function is

$$J(C_i, \Lambda, \mu_i, \nu_i) = \sum (F_i - C_i)^n + \Lambda \left(\sum \frac{2^{C_i/W}}{\alpha_i^2} - P'_{total} \right) - \sum \mu_i C_i + \sum \nu_i (C_i - C_i^0). \tag{3.71}$$

Differentiating with respect to C_i gives

$$\frac{\partial J}{\partial C_i} = -n(F_i - C_i)^{n-1} + \frac{\Lambda'}{\alpha_i^2} 2^{C_i/W} - \mu_i + \nu_i = 0, \tag{3.72}$$

to which we can apply the Kuhn-Tucker condition, to determine in what case we have $C_i = 0$.

When $C_i > 0$, we have $\mu_i = 0$, which leads (3.72) to

$$n(F_i - C_i)^{n-1} = \frac{\Lambda'}{\alpha_i^2} 2^{C_i/W} + \nu_i = \frac{\Lambda'}{\alpha_i^2} \left(1 + \frac{\alpha_i^2 P_i}{W N_0} \right) + \nu_i > \frac{\Lambda'}{\alpha_i^2}, \tag{3.73}$$

which results from $P_i > 0$ and $\nu_i \geq 0$. We notice that this is identical to the previous result, (3.42), where ν_i is not used and $C_i > C_i^0$ (i.e., $P_i > P_0$) is discarded without a loss of optimality.

When $C_j = 0$ ($< C_i^0$), on the other hand, we have $\mu_j \geq 0$ and $\nu_j = 0$, and from (3.72),

$$\mu_j = \frac{\Lambda'}{\alpha_j^2} - n F_j^{n-1} \geq 0 \tag{3.74}$$

$$\Longrightarrow \frac{\Lambda'}{\alpha_j^2} = n F_j^{n-1} + \mu_j \geq n F_j^{n-1}. \tag{3.75}$$

From (3.73) and (3.75) with common Λ', we have

$$n\alpha_i^2 (F_i - C_i)^{n-1} > \Lambda' \geq n\alpha_j^2 F_j^{n-1}, \tag{3.76}$$

which gives

$$\alpha_i^{\frac{2}{n-1}} F_i > \alpha_i^{\frac{2}{n-1}} (F_i - C_i) > \alpha_j^{\frac{2}{n-1}} F_j \quad \text{for } n > 1, \tag{3.77}$$

[6]The rest condition $C_i \leq F_i$ will be satisfied implicitly by a nonnegative Lagrangian multiplier $\Lambda' = \frac{\ln 2}{W} \Lambda$ as in Section 3.2.

where the index i and j represent $\{C_i > 0\}$ and $\{C_j = 0\}$ respectively. Thus, $\alpha_j^{\frac{2}{n-1}} F_j$ should be as small as possible when $C_j = 0$. This proves the argument that we have to provide available active beams for higher attenuation-weighted-demand cells in order to minimize the deviation cost function for $n > 1$. To the K highest attenuation-weighted-demand cells, power is allocated in the same way as described in Section 3.2 by using a proper cost function. Note that with identical signal attenuation for every cell, we should give active beams to K highest demand cells by only considering F_i as shown in Choi and Chan's work [10]. In the presence of different channel fading, as the order of the cost function, n, increases, the difference between the attenuation weights becomes less significant, i.e.,

$$(\alpha_i/\alpha_j)^{\frac{2}{n-1}} \rightarrow 1 \tag{3.78}$$

as $n \rightarrow \infty$ for $\alpha_i > \alpha_j$, and traffic demand alone is dominant. As explained in Section 3.2, higher order cost functions provide better proportional fairness according to traffic demand and take channel conditions into less consideration. On the other hand, lower order cost functions give higher total capacity by making the attenuation weights more important, i.e.,

$$(\alpha_i/\alpha_j)^{\frac{2}{n-1}} \rightarrow \infty \tag{3.79}$$

as $n \rightarrow 1$ for $\alpha_i > \alpha_j$, and focusing more on channel conditions.

We can ask a question of how many active beams K are needed to serve N cells reasonably well. Fig. 3-13 shows the number of active beams required to cover 90% of the total demand, for the same example as in Fig. 3-7 and 3-9. We recognize that a large number of parallel beams are wasted and a modest number can do well enough in many instances. The steeper the slope is, the less beams are required due to less uniformity of traffic. Thus, a reasonable solution is to have a smaller number of active beams and efficiently share them by scheduling among cells. Furthermore, we do not need many modulators with a small number of active beams, so that the system can be more cost-effective and simpler. With this "greedy" policy of serving

K high attenuation-weighted-demand cells, small attenuation-weighted-demand cells may not receive any resources for a long time. So we need to modify the downlink beam scheduling policy, to allocate some capacity for small attenuation-weighted-demand if the delay of a cell is longer than a given deadline.

Assuming very fast antenna beam switching technology with no significant overhead cost, we can time-share a small number of narrow spotbeams efficiently over many cells in the service area. The current satellite switching technique can be as fast as microseconds for power allocation and beam switching by the use of solid state power amplifier to feed about $10 \sim 100$ array elements of a phased array antenna if a priori conditions and variables are provided. In the future, we expect such fast switching technology to be applicable to a much larger scale of antenna (1,000 $\sim$ 100,000 elements and $100 \sim 1{,}000$ beams), in order to support small beams over high frequency bands. As explained in Section 3.1, by locating active beams far enough at each time slot, we can mitigate the interbeam interference problem. Power allocation and beam scheduling should be jointly based on traffic distribution, channel conditions and delay constraints. In practice, we can achieve the time average of capacity required to meet the constraints, by switching spotbeams very quickly over the period of interests.

3.6 Summary

In data satellite communication networks, efficient utilization of the limited amount of precious on-board resources such as power, spotbeam, transmitter and receiver is critical to enhance the system performance to the point of being economically competitive with alternative modalities. To provide responsible satellite downlink services, one should not allocate resources based on only maximizing system capacity but also based on traffic demand and channel conditions. In this chapter, we have shown, for a simplified model and a minimization problem, the optimum solution for satellite downlink multibeam power allocation based on demand and link qualities. We have

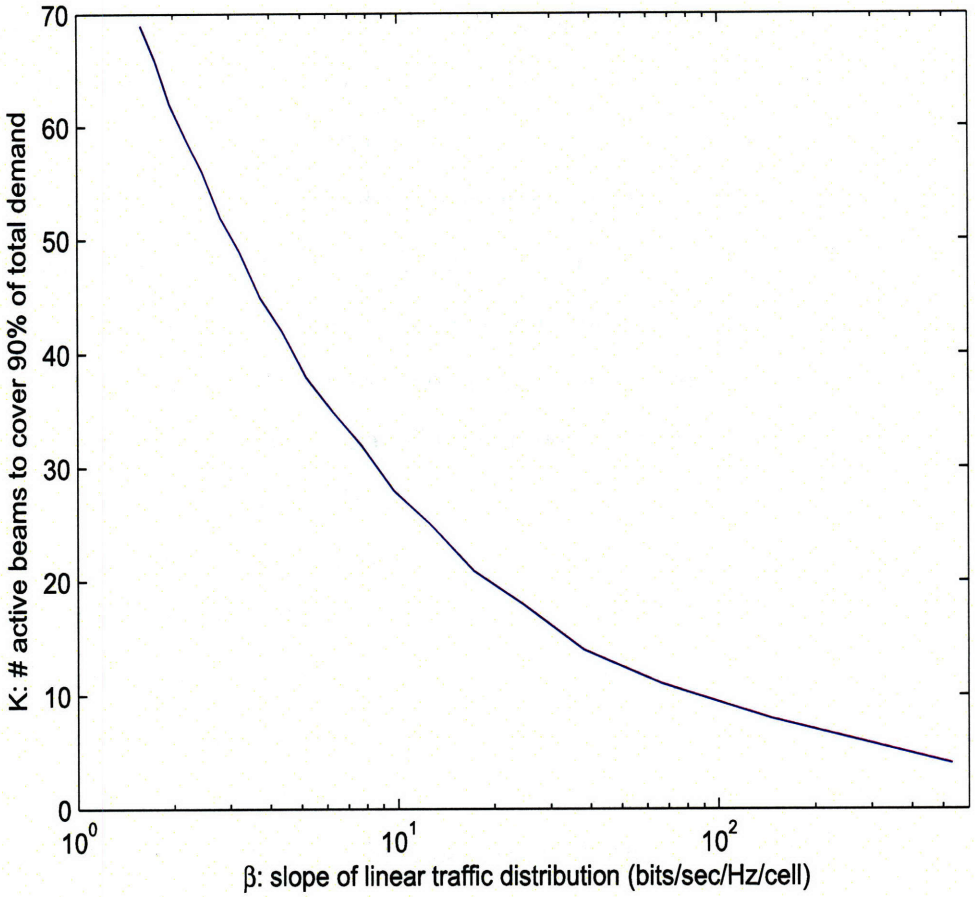

Figure 3-13: Number of active beams required to cover 90% of the total demand, as a function of the slope of linear traffic distribution, to serve $N = 100$ cells

modeled multiple spotbeam downlink capacity. Then, we have found optimum power and spotbeam allocation analytically. Different cost functions for power and beam allocation provide various ways to trade-off between maximizing total capacity and providing proportional fairness. Substantial power gains and fairness advantages can be realized by using parallel multibeams with optimum power allocation. A simple queueing analysis in the steady-state indicates that if a cell has traffic with a tighter delay constraint, more power and capacity should be allocated to give higher priority to this traffic even if the channel is in a poor condition.

Nonuniformity and fluctuation of traffic play important roles in real-life system

performances. Such power allocation and beam scheduling method is vital for a satellite system serving bursty and unscheduled computer data traffic. The more unbalanced traffic is, the larger power gain can be realized and the less number of active beams are required by adaptively providing more (less) power and beam to more (less) demanding cells. When the traffic at each cell increases or backs off, the cells can be scanned sequentially across the coverage area to satisfy the demand. The notion of traffic fluctuation should be also coupled with the dynamic routing and flow/congestion control policy over satellite-terrestrial networks.

3.7 Appendix: Proof of the Optimality of P_i in Section 3.2

Here we prove that the optimality of P_i still holds even after it discards $P_i > P_0$ and $P_i < 0$ in (3.14) and (3.18) respectively by using Theorem 4.4.1 in Gallager's textbook [22].

Theorem *Let $J(x)$ be a convex function of $x = (x_1, ..., x_k)$ over the region $\mathbf{R}$ when x is a probability vector. Assume that the partial derivatives, $\partial J(x)/\partial x_k$ are defined and continuous over the region $\mathbf{R}$ with the possible exception that $\lim_{x_k \to 0} \partial J(x)/\partial x_k$ may be $+\infty$. Then (3.80) and (3.81) are necessary and sufficient conditions on a probability vector x to maximize J over the region $\mathbf{R}$.*

$$\frac{\partial J(x)}{\partial x_k} = \Lambda \quad \textit{all } k \textit{ s.t. } x_k > 0 \tag{3.80}$$

$$\frac{\partial J(x)}{\partial x_k} \leq \Lambda \quad \textit{all } k \textit{ s.t. } x_k = 0 \tag{3.81}$$

where Λ is the same as in the Lagrangian method.

As for $P_i < 0$ first, our function

$$J(P_i) = \sum \left[F_i - \frac{\alpha_i^2 P_i}{N_0 \ln 2} \right]^2 \tag{3.82}$$

in the small SNR region is concave and we want to minimize this over the convex region $P_i \geq 0$. Instead of this, to apply the theorem, we consider maximizing the convex function $-J(P_i)$. When we use all power

$$\sum P_i = P_{total}, \tag{3.83}$$

i.e.,

$$\sum P_i / P_{total} = 1, \tag{3.84}$$

the problem is identical to maximizing a convex function $-J(P_i)$ of a probability vector P_i/P_{total}. From the theorem, the necessary and sufficient conditions for the maximization of $-J(P_i)$ in the low SNR case are

$$-\frac{\partial J}{\partial P_i} = -\left(\frac{\alpha_i^2 P_i}{N_0 \ln 2} - F_i \right) \frac{2\alpha_i^2}{N_0 \ln 2} = \Lambda \quad \text{for } P_i > 0 \tag{3.85}$$

$$-\frac{\partial J}{\partial P_i} = -\left(\frac{\alpha_i^2 P_i}{N_0 \ln 2} - F_i \right) \frac{2\alpha_i^2}{N_0 \ln 2} \leq \Lambda \quad \text{for } P_i = 0, \tag{3.86}$$

which lead to (3.18). This completes the proof that (3.18) is the optimal solution in the small SNR case, though $P_i < 0$ is discarded.

For $P_i > P_0$, with $\Phi_i \equiv P_0 - P_i$, we know that a new function

$$\tilde{J}(\Phi_i) = \sum \left[F_i - W \log \left(1 + \frac{\alpha_i^2 (P_0 - \Phi_i)}{W N_0} \right) \right]^2 \tag{3.87}$$

is still concave over the convex region of $\{\Phi_i\}$. Using the same way as above, we can discard $\Phi_i < 0$, which is indeed $P_i > P_0$, without losing the optimality.

Chapter 4

Joint Multibeam Allocation and Congestion Control for Multiple Beam Antenna

In the previous chapter, we suppressed the issue of controlling backed-up excess traffic and its delay, and focused only on the long-term average gain in terms of Shannon capacity and power efficiency. The most challenging design task to maximize the network efficiency is that the resource allocation and scheduling problem should be considered from the viewpoint of joint optimization over multiple network layers. In this chapter, we couple a multibeam allocation problem with the congestion control of incoming traffic over average delay constraints. Congestion control prevents excessive packet loss and stabilizes the system with an acceptable queueing delay. Here, in particular, we use a form of admission control by allowing only a fraction of incoming traffic, based on the average delay and available resource in the system. We consider the deployment of a multiple beam antenna with a traveling wave tube amplifier (TWTA) for each spotbeam (Fig. 4-1). Channel conditions are assumed to be quasi-static over the period of interest (seconds) and beam switching is assumed to be very fast by the use of advanced electronic or electro-optical beam switching technologies.

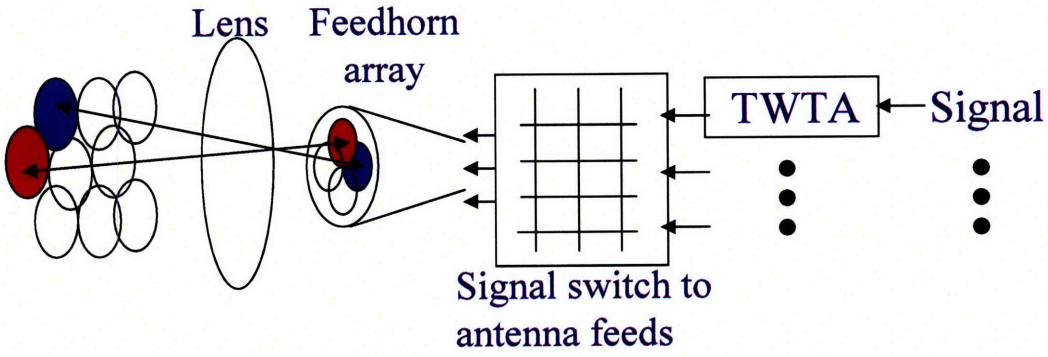

Figure 4-1: A schematic of multiple beam antenna

In Section 4.1, we formulate the throughput maximization problem by controlling admission of incoming traffic with stability, average delay, and beam-sharing constraints. In Section 4.2, an analytic solution for joint beam allocation and congestion control is obtained by using queueing theory. In Section 4.3, we compare this jointly optimized scheme with the uniform beam allocation scheme with respect to throughput, queueing delay, and fairness. Some examples are given in Section 4.4 to highlight the impact of changes of traffic demand and channel conditions to the performance. Section 4.5 summarizes this chapter.

4.1 Formulation

We want to allocate efficiently a limited amount of on-board transmission power and a small number of K active beams among many small N ($> K$) cells within a satellite coverage area. We also want to maximize the throughput of the system with reasonable queueing delays. For this purpose, we introduce a congestion control back-off parameter θ ($0 \leq \theta \leq 1$), which adjusts the amount of incoming traffic to all the queues uniformly, based on channel conditions and average delay constraints. A single back-off parameter is used for all users to achieve proportional fairness. We consider a multibeam satellite in a steady state where incoming traffic of the average rate A_i (without congestion control) is presented to the i^{th} cell with channel capacity

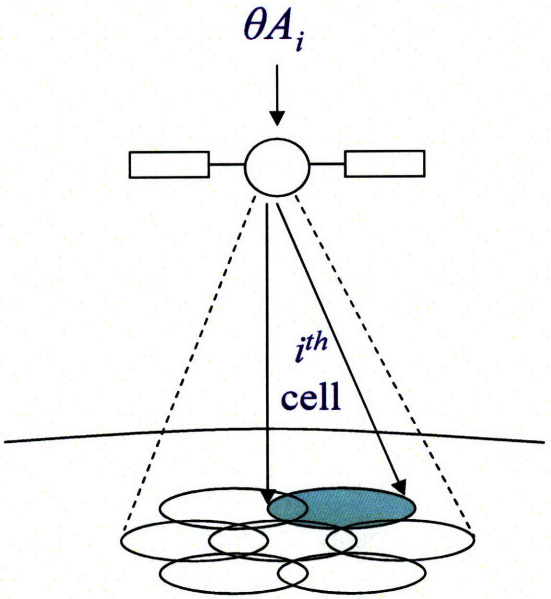

Figure 4-2: A multibeam downlink satellite with congestion-controlled incoming traffic

C_i if the full power of a TWTA is being used for transmission (Fig. 4-2). With the idealized assumption of infinite buffer size (i.e., no packet loss due to a full queue), any type of general traffic pattern is acceptable in this case.[1] If A_i is too large for the system capacity, congestion control is triggered and incoming traffic should be backed off to θA_i with the maximum θ that can stabilize the system with an upper-bounded average and a finite variance for the amount of accumulated traffic in the queue. By maximizing θ, we can maximize the throughput of the system with fairness since θ is universal to all users.

Channel conditions are assumed to be quasi-static with constant signal attenuation over the period of interest. This assumption is reasonable because the packet processing time and transmission deadlines are in general much shorter than the coherence time of signal attenuation due to rain, which is of the order of minutes or hours. We assume the use of TWTAs and a multiple beam antenna, which has a

[1]We should also consider the case when traffic is unpredictable and bursty. θ should be time-varying in this case, which is covered at the end of Chapter 5. However, in this chapter θ is confined to be fixed for the interval of interest.

power constraint of $P_i \leq P_0$ for each beam, where P_i is the power allocated to the i^{th} cell. We assume that the TWTAs are driven well into saturation for efficiency and thus frequency multiplexing to provide multiple beams is not viable. In such a situation, it is optimal to use full power of $P_i = P_0$ for all active K beams all the time over quasi-static channel conditions, and thus, to achieve full channel capacities, which results in constant channel capacity C_i for each channel over the time interval of interest. (Of course, C_i can differ from beam to beam due to local weather conditions.)

For given A_i and C_i,[2] we seek the necessary condition in which the system is stable, i.e., the average queueing delay is finite. We use a binary variable $z_i(t)$ to indicate whether the i^{th} cell is served (1) or not (0) at discrete time $t = 0, 1, 2, ...$ And for the time interval $[0, T]$ in a steady state, we define

$$\zeta_i = \sum_{t=0}^{T} z_i(t), \tag{4.1}$$

which represents how long the cell is serviced during $[0, T]$. Then, for a stable system, the whole incoming traffic during the time interval should be less than the total capacity allocated, i.e.,

$$\theta A_i \cdot T \leq \zeta_i \cdot C_i \quad \Longrightarrow \quad \theta \cdot \frac{A_i}{C_i} \leq \frac{\zeta_i}{T}. \tag{4.2}$$

By summing both sides over i, we obtain

$$\theta \cdot \sum_{i=1}^{N} \frac{A_i}{C_i} \leq \sum_{i=1}^{N} \frac{\zeta_i}{T} \leq \frac{KT}{T} = K, \tag{4.3}$$

where we have

$$\sum_{i=1}^{N} \zeta_i \leq KT \tag{4.4}$$

[2]As for A_i, in general, we are more interested in packets/sec than bits/sec. Since C_i is given in terms of bits/sec, we should compare A_i with $C_i/\bar{l_p}$ (or $A_i \cdot \bar{l_p}$ with C_i), where $\bar{l_p}$ is the average packet length in terms of bits. For simplicity of the analysis, we assume that $\bar{l_p} = 1$ here. Since our result is mainly a steady-state average, we can simply replace C_i with $C_i/\bar{l_p}$ to get the result in terms of packets (especially with the $M/M/1$ assumption) or A_i with $A_i \cdot \bar{l_p}$ in terms of bits.

since we are limited to have K beams. Thus, we conclude that it is necessary to satisfy

$$\theta \sum_{i=1}^{N} \frac{A_i}{C_i} \leq K \tag{4.5}$$

for the system to be stable. The same result is also shown in Neely *et al.*'s work [44] without a congestion control parameter.

In addition, even with the upper-bounded average queueing delay (which is easily related to the average amount of accumulated traffic by Little's theorem), the queue can be unstable with an infinite variance σ_i^2 for the amount of accumulated traffic F_i of the i^{th} queue. Thus, we must also have $\sigma_i^2 < \infty$ for every i.

The problem[3] can be formulated as

$$\text{maximize } \theta \tag{4.7}$$

$$\text{subject to } \quad 0 \leq \theta \leq 1 \tag{4.8}$$

$$\theta \cdot \sum_{i=1}^{N} \frac{A_i}{C_i} \leq K \tag{4.9}$$

$$\theta \cdot \frac{A_i}{C_i} \leq 1 \quad \text{for every } i = 1, ..., N \tag{4.10}$$

$$\sigma_i^2 < \infty \quad \text{for every } i = 1, ..., N \tag{4.11}$$

$$\bar{d}_i \leq \Delta_i \quad \text{for every } i = 1, ..., N \tag{4.12}$$

$$\text{and} \quad \sum_{i=1}^{N} z_i(t) \leq K \quad (z_i(t) = 0 \text{ or } 1 \text{ for every } t \text{ and } i), \tag{4.13}$$

where $\bar{d}_i$ is the average queueing delay of the i^{th} queue, and Δ_i (> 0) is a given delay deadline and may be different from beam to beam.

[3]In general, a resource allocation problem is given as

$$\text{maximize} \quad \sum_i f_i(P_i) \tag{4.6}$$

with power/capacity constraints and others as needed, where the general utility function f_i is defined to be concave with respect to power and capacity. In this thesis, even with a simple linear back-off parameter θ, we consider average delay constraints. The concave delay penalty function will be incorporated in the cost function when we develop a dynamic algorithm in Chapter 5.

We want to maximize the back-off parameter θ of incoming traffic A_i while assuring system stability with conditions (4.9), (4.10) and (4.11). Condition (4.9) is for stability of the whole system with K active beams and conditions (4.10) and (4.11) are for each queue. Condition (4.12) represents a constraint of average delay $\bar{d}_i$ within a given target deadline Δ_i. Note that $\bar{d}_i$ is a function of $\theta A_i, C_i$, and $z_i(t)$. Throughout the thesis, we consider only an average delay constraint, not a hard deadline constraint for each packet, i.e., $d_i(t) \leq \Delta_i$. Time-varying traffic demand and channel conditions make it infeasible to apply the hard deadline constraint all the time because deep fading events even for a short duration, which happen in reality from time to time, can prohibit any utilization of resource and lead to violation of the constraint. For the same reason, we consider the quasi-static average channel conditions in the analysis. One may consider outage, which gives the probability that the delay (or any quality of service in general) exceeds the given threshold, but this is beyond the scope of the thesis. Condition (4.13) is for the transmitter-sharing constraint. Intuitively, shorter delay constraints require smaller θ to have a smaller number of packets in the queues and/or larger $\bar{z}_i$ to secure more service for the cell.

Fig. 4-3 shows the block diagram of this beam allocation and congestion control scheme. Perfect information on channels and the system is assumed for decision-making. In practice, the information can be inferred based on estimates on a reciprocal channel or from direct feedback in a return channel. In [8, 9], it is shown that signal attenuation due to rain can be estimated with good accuracy (within a 1 dB error in 4 seconds ahead) based on a simple one-pole model. Thus, even in the presence of long propagation delays over satellite-ground links, beam allocation and congestion control parameters can be determined in advance, for this scheme to operate well.

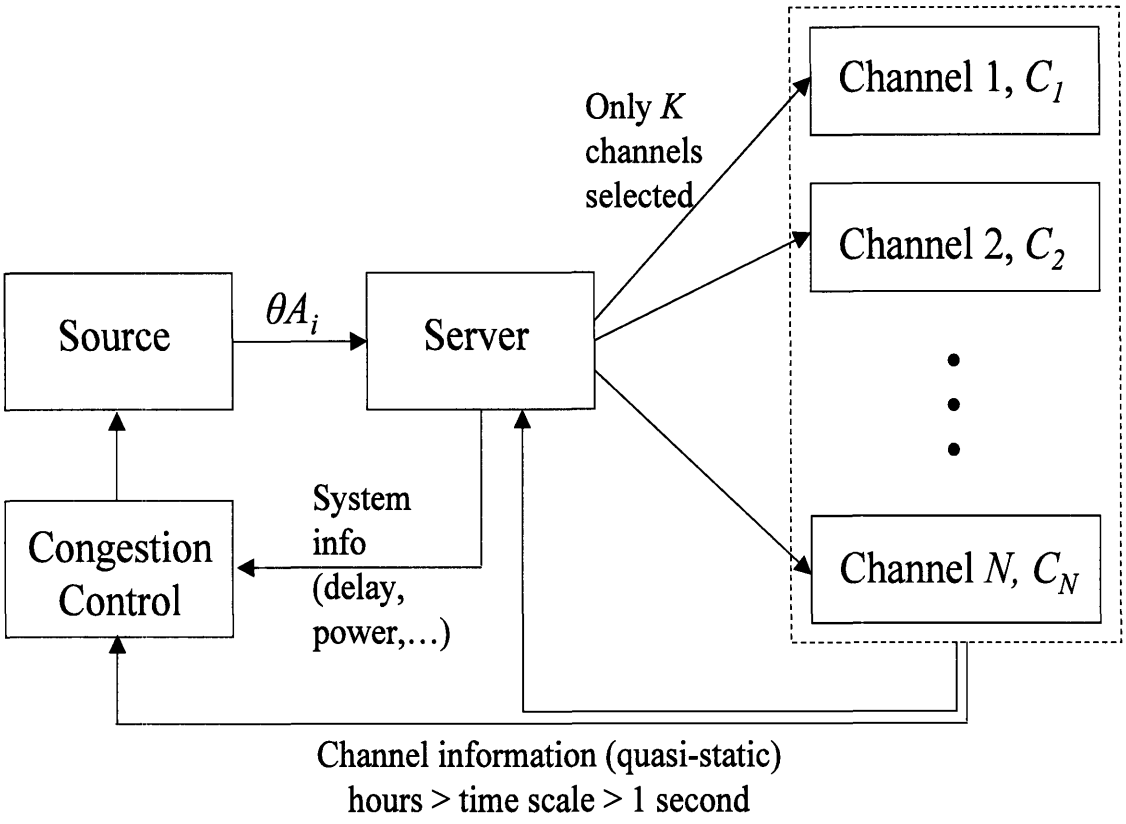

Figure 4-3: Block diagram of the beam allocation and congestion control scheme

4.2 Analysis

Before solving the problem (4.7), we will simplify some constraints. First, the stability conditions of (4.9) and (4.10) are redundant since the finite deadline constraint (4.12) guarantees the finite amount of accumulated traffic in the system. Next, since a binary variable $z_i(t)$ makes the problem complicated, we assume that the average delay depends only on the steady-state average of $z_i(t)$, $\bar{z}_i$ (for example, the $M/M/1$ queue), so that we can remove the binary condition from the formulation. This approximation is good when the active beams cycle through the cells more rapidly than the time scale of the deadlines. When we assume that beam switching is very fast with no significant overhead cost, we can achieve $\bar{z}_i$ by time-sharing beams and changing $z_i(t)$ properly. The current switching technique can be as fast as milliseconds

to feed about dozens of active beams of a multiple beam antenna over microwave bands if a priori conditions are provided. The faster switching technology is expected for a much larger scale of antenna (100 $\sim$ 1,000 beams) in the future, and the use of phased array antenna and solid state power amplifiers can be an alternative choice, which is studied in Chapter 5.

Thus, the finite variance, deadline and transmitter-sharing constraints in terms of θ and $\bar{z}_i$ are enough to describe the problem, given as

$$\text{maximize } \theta \tag{4.14}$$

$$\text{subject to} \quad 0 \leq \theta \leq 1 \tag{4.15}$$

$$\sigma_i^2 < \infty \quad \text{for every } i = 1, ..., N \tag{4.16}$$

$$\bar{d}_i \leq \Delta_i \quad \text{for every } i = 1, ..., N \tag{4.17}$$

$$\text{and} \quad \sum_{i=1}^{N} \bar{z}_i \leq K \quad \text{with } 0 \leq \bar{z}_i \leq 1 \text{ for every } i. \tag{4.18}$$

In general, it is hard to account for the finite variance constraint (4.16). However, for the example of an $M/M/1$ queue, it can be done analytically. If we set aside condition (4.16) for the time being, the solution can be obtained in a straightforward way. We see that θ and $\bar{z}_i$ are coupled in the deadline condition (4.17). Let us define

$$\bar{d}_i \equiv f_i(\theta, \bar{z}_i). \tag{4.19}$$

By nature of the delay function, f_i is an increasing function of θ (more incoming traffic increases the delay) and decreasing function of $\bar{z}_i$ (more service decreases the delay). If we assume that there exists an inverse function for $\bar{z}_i$, $f_i^{-1}(\Delta_i; \theta)$ with given θ, we have

$$\bar{z}_i \geq f_i^{-1}(\Delta_i; \theta) \tag{4.20}$$

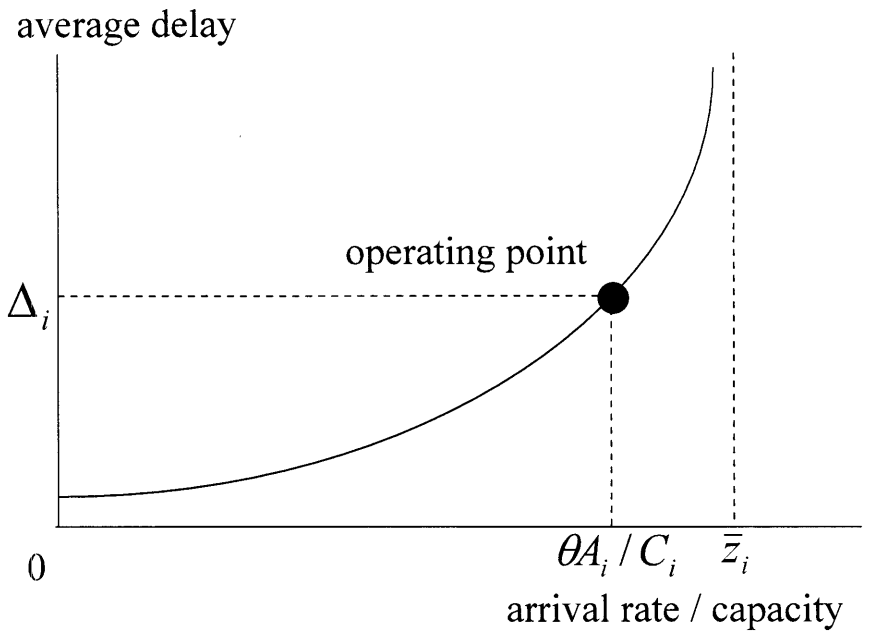

Figure 4-4: A delay-arrival rate/capacity curve for the i^{th} queue

from (4.17). By combining this with condition (4.18), we have

$$f_i^{-1}(\Delta_i;\theta) \leq 1 \quad \text{and} \quad \sum_{i=1}^{N} f_i^{-1}(\Delta_i;\theta) \leq K, \tag{4.21}$$

and we can determine the optimum θ. $\bar{z}_i$ is given by (4.20) to satisfy $\bar{z}_i \leq 1$ and $\sum_i \bar{z}_i \leq K$.

For a given average delay deadline Δ_i and congestion-controlled incoming traffic θA_i, we can determine the operating point over a curve relating average delay and the ratio of arrival rate and capacity (Fig. 4-4). By extending this curve toward an infinite delay, we obtain the asymptote of the arrival rate over capacity, which is equal to average beam allocation $\bar{z}_i$ required for serving incoming traffic θA_i at the delay of Δ_i. The average capacity of the i^{th} cell can be achieved by adjusting $z_i(t)$ and is equal to $\bar{z}_i C_i$, where C_i is assumed to be fixed. We note that the utilization factor of $\frac{\theta A_i}{\bar{z}_i C_i}$ is strictly less than one due to the finite average delay requirement.

To get a meaningful insight for the analysis, we now simply assume that each queue of the satellite downlink beams resembles an $M/M/1$ queue. Suppose a Poisson arrival

process of incoming traffic with average rate θA_i and an exponentially distributed traffic packet size with average transmission rate $\bar{z}_i C_i$, and then we find an optimal θ and the corresponding $\bar{z}_i$. We have an average delay (per bit) of $M/M/1$ queue [4], given as

$$\bar{d}_i = f_i(\theta, \bar{z}_i) = \frac{1}{\bar{z}_i C_i - \theta A_i} \leq \Delta_i, \tag{4.22}$$

which leads to

$$1 \geq \bar{z}_i \geq \theta \cdot \frac{A_i}{C_i} + \frac{1}{C_i \Delta_i} \tag{4.23}$$

and

$$K \geq \theta \cdot \sum_{i=1}^{N} \frac{A_i}{C_i} + \sum_{i=1}^{N} \frac{1}{C_i \Delta_i}. \tag{4.24}$$

Then, the optimum θ_{joint} for the joint beam allocation and congestion control problem is given as

$$\theta_{joint} = \min \left\{ \frac{1 - \frac{1}{C_i \Delta_i}}{\frac{A_i}{C_i}}, \frac{\frac{K}{N} - \frac{1}{N} \sum_i \frac{1}{C_i \Delta_i}}{\frac{1}{N} \sum_i \frac{A_i}{C_i}}, 1 \right\} \tag{4.25}$$

and the corresponding $\bar{z}_i$ is given as

$$\begin{aligned} \bar{z}_i &\geq \theta_{joint} \cdot \frac{A_i}{C_i} + \frac{1}{C_i \Delta_i}, \\ &= \frac{A_i}{C_i} \cdot \min \left\{ \frac{1 - \frac{1}{C_j \Delta_j}}{\frac{A_j}{C_j}}, \frac{\frac{K}{N} - \frac{1}{N} \sum_j \frac{1}{C_j \Delta_j}}{\frac{1}{N} \sum_j \frac{A_j}{C_j}}, 1 \right\} + \frac{1}{C_i \Delta_i}, \end{aligned} \tag{4.26}$$

where the equality holds when $\theta_{joint} < 1$ with $\bar{d}_i = \Delta_i$. When $\theta_{joint} = 1$, there is no congestion control needed and more $\bar{z}_i$ can be allocated with $\bar{d}_i < \Delta_i$.

In (4.23) and (4.24) stability conditions for each queue and the whole system are included as described before. Moreover, since we have a finite delay deadline Δ_i, the additional price $\frac{1}{C_i \Delta_i}$ is imposed for each queue and $\sum \frac{1}{C_i \Delta_i}$ for the whole system. The smaller the deadlines and/or capacities are, the higher the price is, which is reasonable because more urgent services and/or worse channel conditions require more beam allocation. In (4.25) there can be no feasible solution if $\frac{1}{C_i \Delta_i} > 1$ or $\sum \frac{1}{C_i \Delta_i} > K$,

where the channel condition is too bad to support the required deadline constraint.

As for the finite variance constraint (4.16) of σ_i^2 for accumulated traffic F_i of the i^{th} queue, we can show that $\sigma_i^2 < \infty$ with the utilization factor of

$$u_i \equiv \frac{\theta A_i}{\bar{z}_i C_i}, \tag{4.27}$$

given as

$$\begin{aligned}
\sigma_i^2 &= E[F_i^2] - \{E[F_i]\}^2 \\
&= \sum_{n=0}^{\infty} n^2 \Pr[F_i = n] - \{E[F_i]\}^2 \\
&= \sum n(n-1)u_i^n(1-u_i) + E[F_i] - \{E[F_i]\}^2 \\
&= u_i^2(1-u_i)\sum n(n-1)u_i^{n-2} + \frac{u_i}{1-u_i} - \left(\frac{u_i}{1-u_i}\right)^2 \\
&= u_i^2(1-u_i)\frac{\partial^2}{\partial u_i^2}\left(\frac{1}{1-u_i}\right) + \frac{u_i}{1-u_i} - \left(\frac{u_i}{1-u_i}\right)^2 \\
&= u_i^2(1-u_i)\frac{2}{(1-u_i)^3} + \frac{u_i}{1-u_i} - \left(\frac{u_i}{1-u_i}\right)^2 \\
&= \frac{u_i}{(1-u_i)^2} < \infty,
\end{aligned} \tag{4.28}$$

where some $M/M/1$ queue results are used such as

$$\Pr[F_i = n] = u_i^n(1-u_i) \tag{4.29}$$

and

$$E[F_i] = \frac{u_i}{1-u_i}. \tag{4.30}$$

$E[\cdot]$ is an ensemble average operator and $\Pr[\cdot]$ represents the probability of the event inside the bracket. As explained before, u_i is strictly less than 1 as a consequence of the finite delay requirement, so that the variance is finite.

We now compare this scheme of joint beam allocation and congestion control with

the simple scheme of uniform beam allocation of

$$\bar{z}_i = \frac{K}{N}. \tag{4.31}$$

With uniform beam allocation, we do not need the transmitter-sharing constraint, but only the deadline constraint of

$$\bar{d}_i = \frac{1}{\frac{K}{N}C_i - \theta A_i} \leq \Delta_i, \tag{4.32}$$

and then the optimum $\theta_{uniform}$ in this case is given as

$$\theta_{uniform} = \min\left\{ \frac{\frac{K}{N} - \frac{1}{C_i\Delta_i}}{\frac{A_i}{C_i}}, 1 \right\}, \tag{4.33}$$

which satisfies

$$\theta_{uniform}\frac{A_i}{C_i} + \frac{1}{C_i\Delta_i} \leq \frac{K}{N}. \tag{4.34}$$

On the other hand, from (4.25) the optimum θ_{joint} with joint beam allocation and congestion control satisfies

$$\theta_{joint}\frac{1}{N}\sum_{i=1}^{N}\frac{A_i}{C_i} + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{C_i\Delta_i} \leq \frac{K}{N} \tag{4.35}$$

and

$$\theta_{joint}\frac{A_i}{C_i} + \frac{1}{C_i\Delta_i} \leq 1. \tag{4.36}$$

In Fig. 4-5, we compare θ_{joint} and $\theta_{uniform}$ by plotting (4.34), (4.35), and (4.36) on the plane of $(\frac{A_i}{C_i}, \frac{1}{C_i\Delta_i})$ and considering the absolute values of the slopes. $\theta_{uniform}$ represents the least steep slope among the lines connecting $(0, \frac{K}{N})$ and $(\frac{A_i}{C_i}, \frac{1}{C_i\Delta_i})$ for every i. In the similar way, θ_{joint} is the minimum absolute value of the slope from $(0, \frac{K}{N})$ to $(\frac{1}{N}\sum\frac{A_i}{C_i}, \frac{1}{N}\sum\frac{1}{C_i\Delta_i})$ or from (0, 1) to $(\frac{A_i}{C_i}, \frac{1}{C_i\Delta_i})$ for every i. Roughly speaking, between (4.34) and (4.35), since $(\frac{1}{N}\sum\frac{A_i}{C_i}, \frac{1}{N}\sum\frac{1}{C_i\Delta_i})$ is located inside the polygon consisting of the points of $(\frac{A_i}{C_i}, \frac{1}{C_i\Delta_i})$, θ_{joint} should be larger than $\theta_{uniform}$. Between

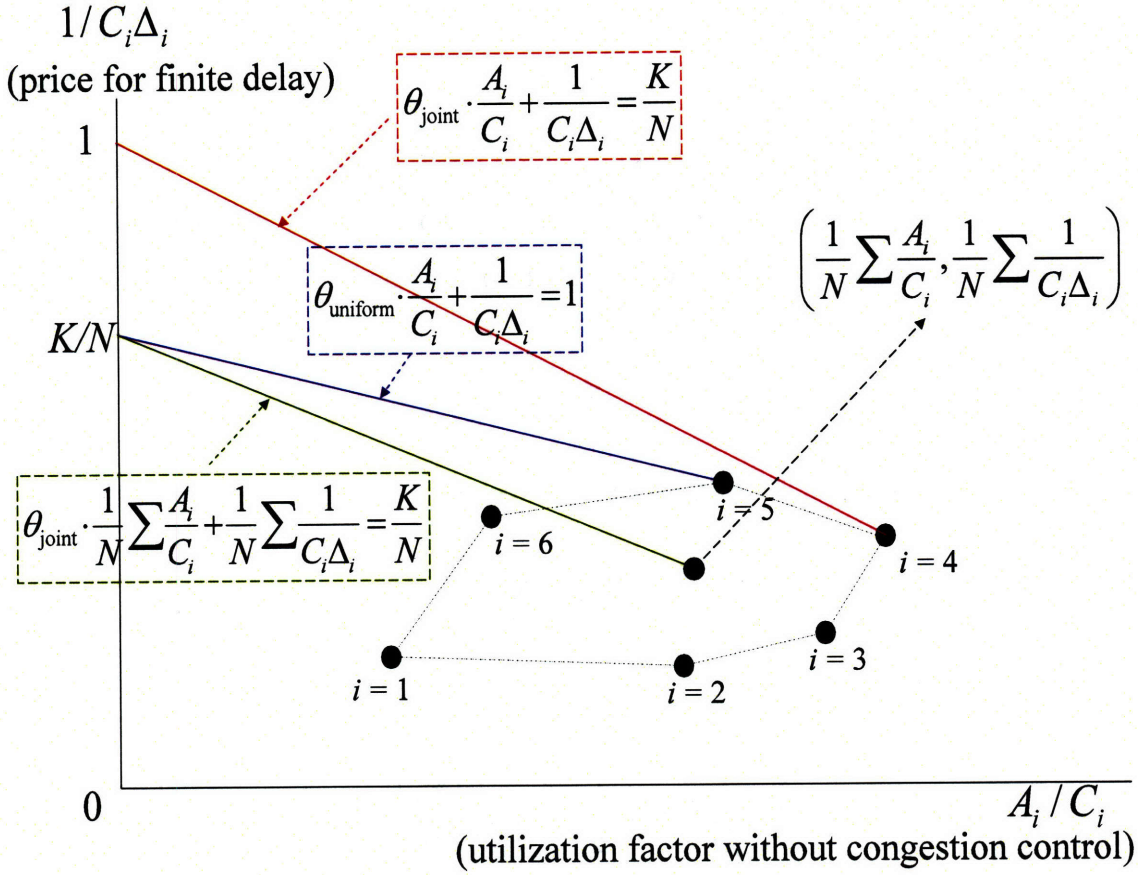

Figure 4-5: Illustration of comparing $\theta_{uniform}$ and θ_{joint} in (4.34; blue), (4.35; green), and (4.36; red)

(4.34) and (4.36), since (0, 1) is located higher than $(0, \frac{K}{N})$, again θ_{joint} should be larger. Therefore, we can show that the joint scheme of beam allocation and congestion control gives larger θ_{joint}, that is, more accepted incoming traffic into the system under the same deadline constraint, compared with the uniform allocation scheme.

4.3 Numerical Results and Discussion

In this section, we provide numerical results under simple scenarios of linearly distributed incoming traffic A_i and channel capacities C_i. In Fig. 4-6, 4-7, and 4-8, for $N = 100$ cells and $K = 20$ active beams, we consider an example of linearly

distributed incoming traffic of

$$A_i = i \cdot \beta, \tag{4.37}$$

where β is a parametric slope. We use identical $C_i \equiv C$ and $\Delta_i \equiv \Delta$ for every i. As the traffic of the most crowded cell A_{100} changes with β, we can compare joint beam allocation and congestion control with uniform beam allocation. Fig. 4-6 verifies

$$\theta_{joint} \geq \theta_{uniform} \tag{4.38}$$

for every traffic distribution. In particular, since

$$\frac{A_N}{C_N} = 2 \cdot \frac{1}{N} \sum \frac{A_i}{C_i} \tag{4.39}$$

for linearly distributed traffic, when $\theta_{uniform} < 1$ and $\theta_{joint} < 1$, we have

$$\theta_{uniform} = \frac{\frac{K}{N} - \frac{1}{C_N \Delta_N}}{\frac{A_N}{C_N}} = \frac{1}{2} \cdot \frac{\frac{K}{N} - \frac{1}{N} \sum \frac{1}{C_i \Delta_i}}{\frac{1}{N} \sum \frac{A_i}{C_i}} = \frac{1}{2} \cdot \theta_{joint}, \tag{4.40}$$

where

$$\frac{1}{N} \sum \frac{1}{C_i \Delta_i} = \frac{1}{C_N \Delta_N} = \frac{1}{C\Delta}. \tag{4.41}$$

Thus, the joint scheme accepts traffic twice as much as the uniform scheme for linearly distributed traffic with congestion control on. For the same reason, the slope A_{100}/C with which θ becomes less than 1 (i.e., congestion control gets triggered) for the joint scheme is a factor of 2 away from that for the uniform scheme ($A_{100}/C = 0.3$ for θ_{joint} while $A_{100}/C = 0.15$ for $\theta_{uniform}$).

Fig. 4-7 compares the corresponding delays (normalized by Δ) of the $N^{th}(= 100^{th})$ cell. When $\theta_{joint} = 1$ (i.e., $A_{100}/C = 0.3$), the delay of the joint scheme is smaller than that of the uniform scheme by allocating more beam to heavier traffic. When $\theta_{uniform} < 1$, since only the N^{th} cell is dominant for deciding $\theta_{uniform}$, other cells are under-utilized below capacities with small $\theta_{uniform}$. On the other hand, when $\theta_{joint} < 1$, the whole system decides θ_{joint} by $\frac{1}{N} \sum \frac{1}{C_i \Delta_i}$ and $\frac{1}{N} \sum \frac{A_i}{C_i}$. All cells are

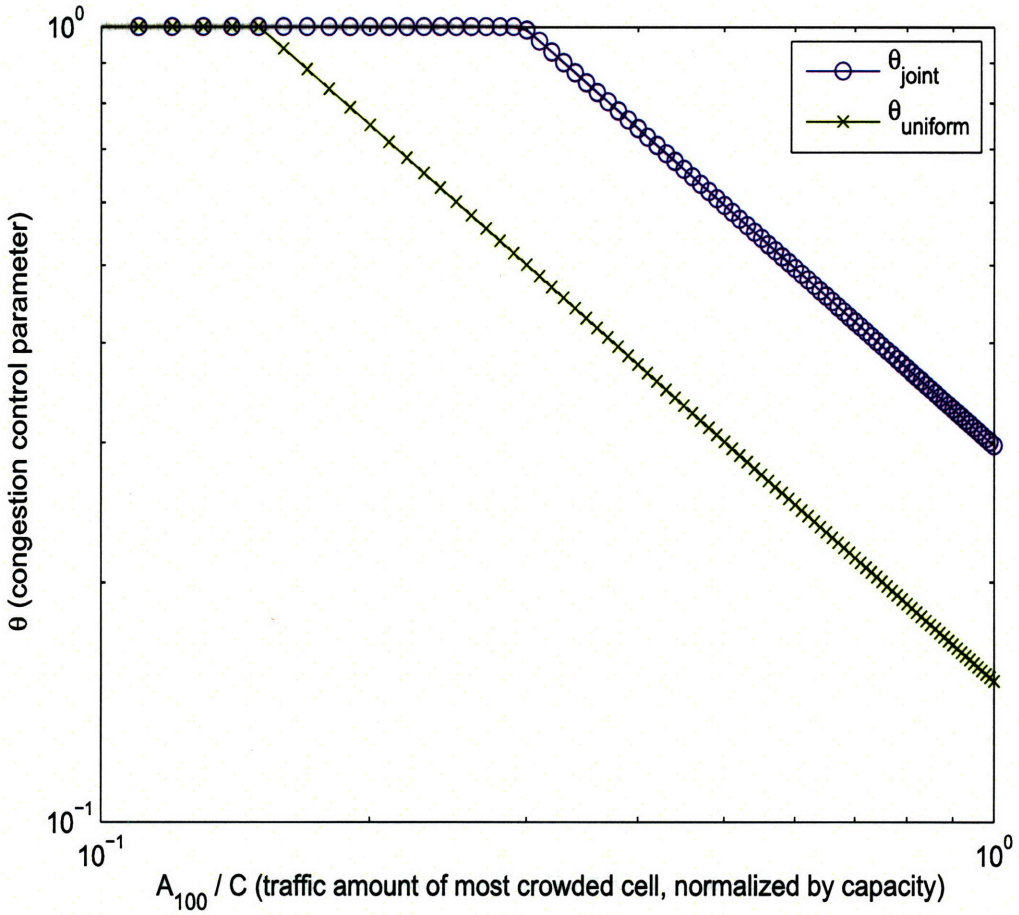

Figure 4-6: Comparing congestion control parameters θ_{joint} and $\theta_{uniform}$ of joint and uniform beam allocation/congestion control as a function of the traffic of the most crowded cell A_{100} (normalized by channel capacity C) for linearly distributed incoming traffic

equally and maximally utilized under the deadline constraint. Fig. 4-8 compares the time fractions of beam allocation for the 1^{st} and 100^{th} $(= N^{th})$ cell, i.e., $\bar{z}_1$ and $\bar{z}_{100}$ for those of uniform beam allocation $\bar{z}_i = K/N$. Before congestion control is on $(A_{100}/C = 0.3)$, as the slope of traffic distribution gets steeper, more beams are allocated to heavier traffic cells. For $\theta_{joint} < 1$, we see that incoming traffic $\theta_{joint}A_i$ is fixed by controlling θ_{joint}, and thus, delay $\bar{d}_i$ and beam allocation $\bar{z}_i$ are also fixed. (Moreover, $\bar{d}_i = \Delta$ for every i.)

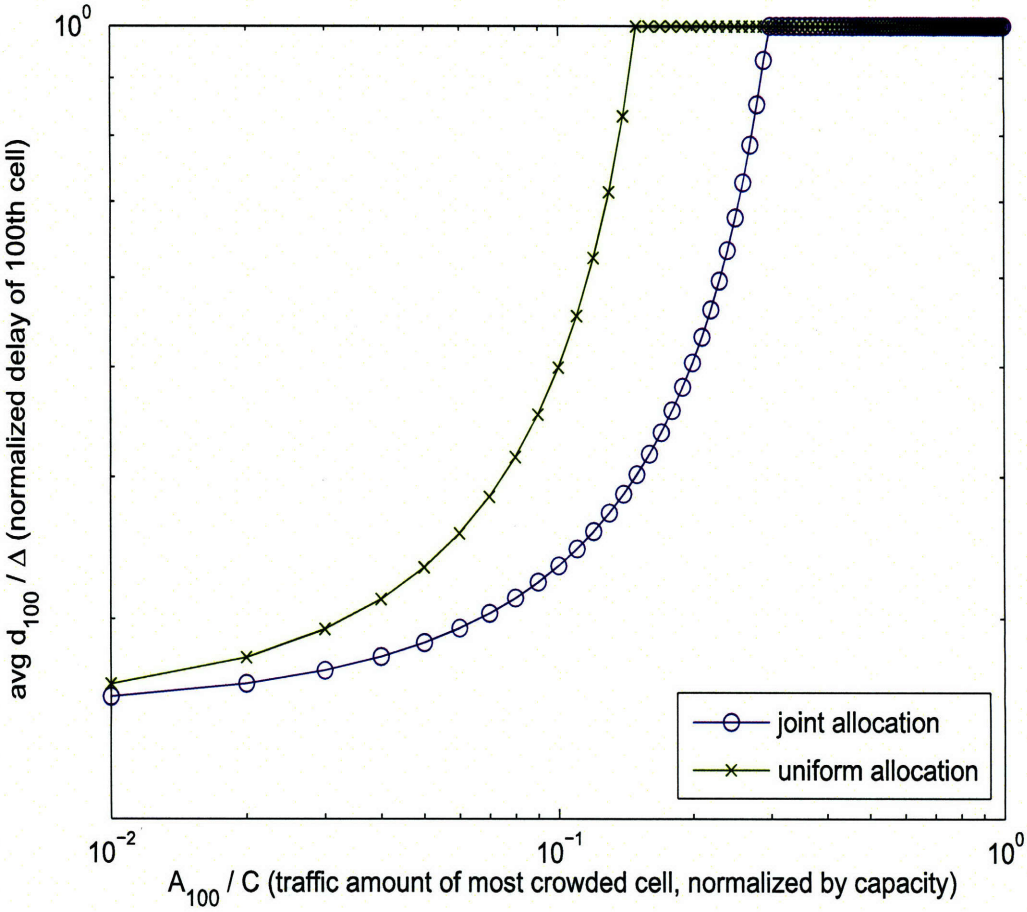

Figure 4-7: Comparing average delays of the $N^{th}(= 100^{th})$ cell normalized by the deadline of joint and uniform beam allocation/congestion control as a function of the traffic of the most crowded cell A_{100} (normalized by channel capacity C) for linearly distributed incoming traffic

In Fig. 4-9, 4-10, and 4-11, we consider linearly distributed capacities of

$$C_i = C_1 - \phi \cdot (i-1), \tag{4.42}$$

where ϕ (> 0) is a parametric slope and C_1 is the fixed capacity of the first cell. Here we use identical $A_i \equiv A$ and $\Delta_i \equiv \Delta$ for every i. We then change the worst channel capacity C_{100} with ϕ and compare two schemes. Again, the joint scheme provides at least one of two advantages: more accepted traffic or a smaller average delay. We note

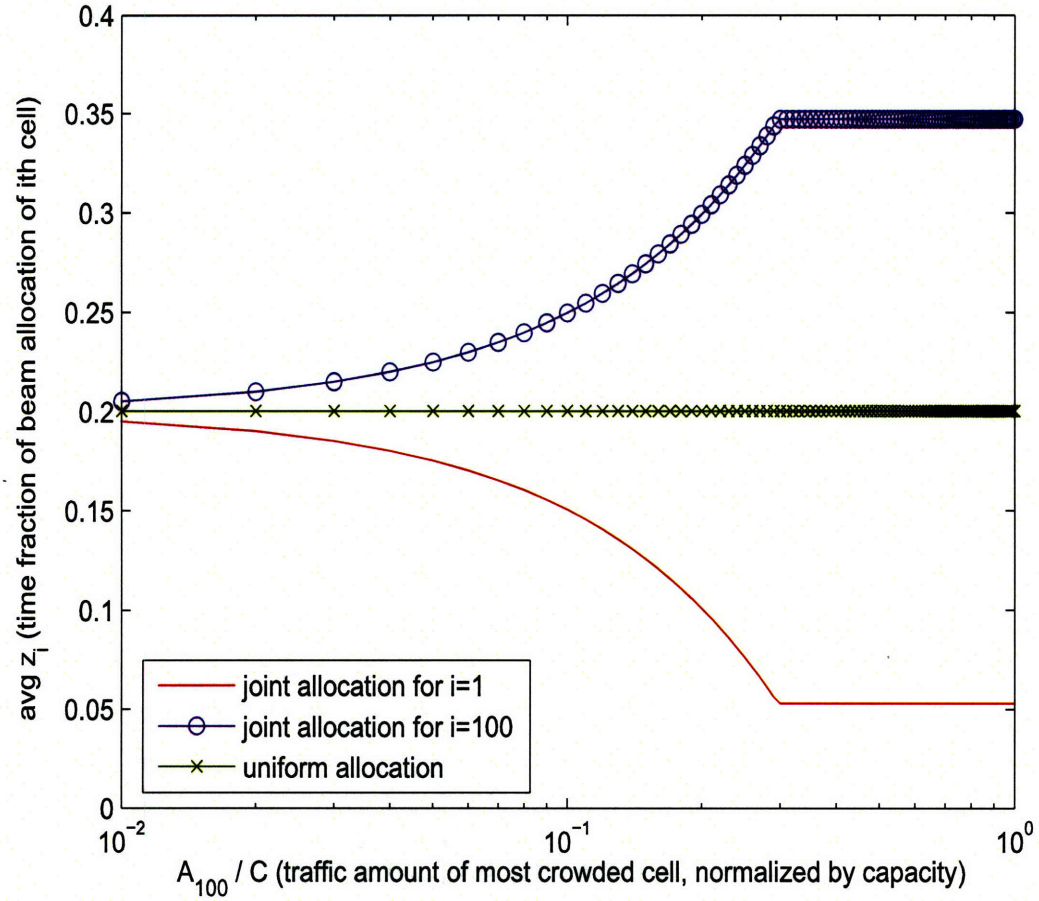

Figure 4-8: Comparing average time fractions of beam allocation of the 1^{st} and $100^{th}(= N^{th})$ cells of joint and uniform beam allocation/congestion control as a function of the traffic of the most crowded cell A_{100} (normalized by channel capacity C) for linearly distributed incoming traffic

that as the channel conditions become severely worse ($C_{100}/C_1 < 0.07$), the worst cell of $i = N$ dominates other cells and gives

$$\theta_{joint} = \frac{1 - \frac{1}{C_N \Delta_N}}{\frac{A_N}{C_N}} \tag{4.43}$$

in Fig. 4-9, and that cell exclusively consumes one beam with $\bar{z}_N = 1$ in Fig. 4-11. On the other hand, as the channel conditions become uniform ($\phi \to 0$), beam allocation converges to the uniform case of $\bar{z}_i = K/N$ for every i (Fig. 4-11).

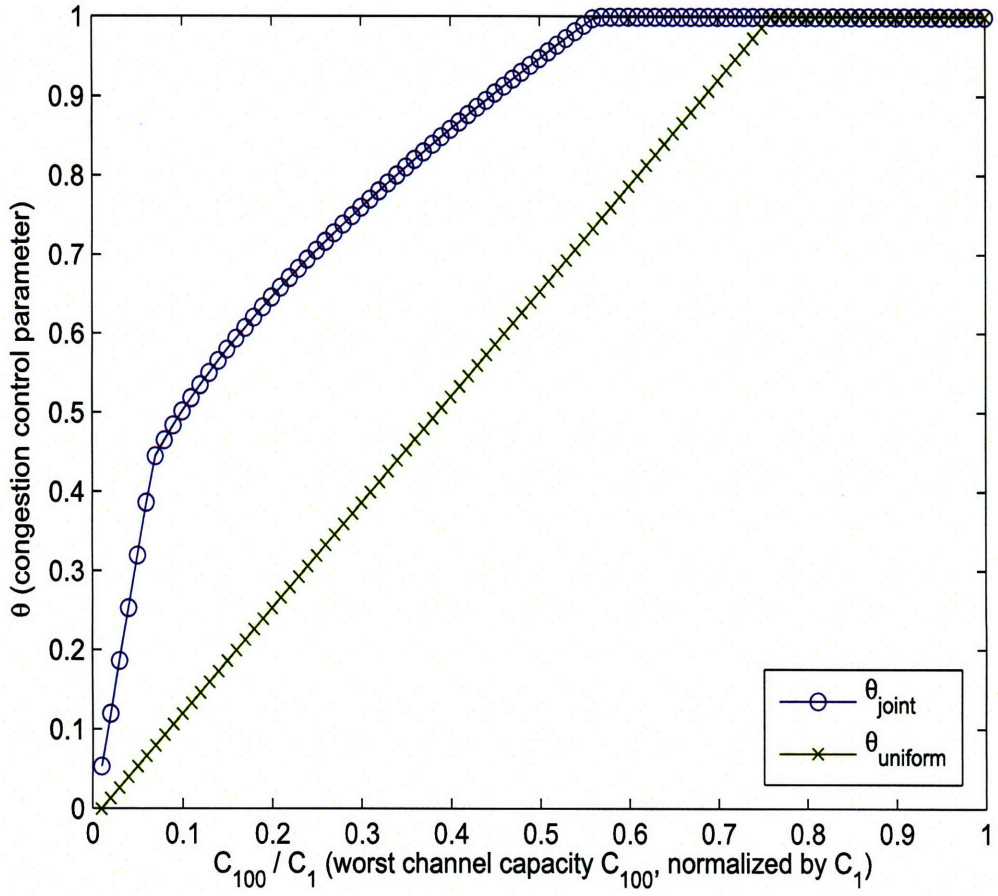

Figure 4-9: Comparing congestion control parameters θ_{joint} and $\theta_{uniform}$ of joint and uniform beam allocation/congestion control as a function of the worst channel capacity C_{100} (normalized by the fixed capacity of the 1^{st} cell, C_1) for linearly distributed channel capacities

In summary, the joint scheme outperforms the uniform scheme by having a smaller average delay for the most crowded and critical cell when congestion control is off, and accepting more incoming traffic and thus better utilizing the capacities under deadline constraints when congestion control is on. Moreover, since the joint scheme allocates spotbeams based on traffic demand, better fairness amongst users can be assured.

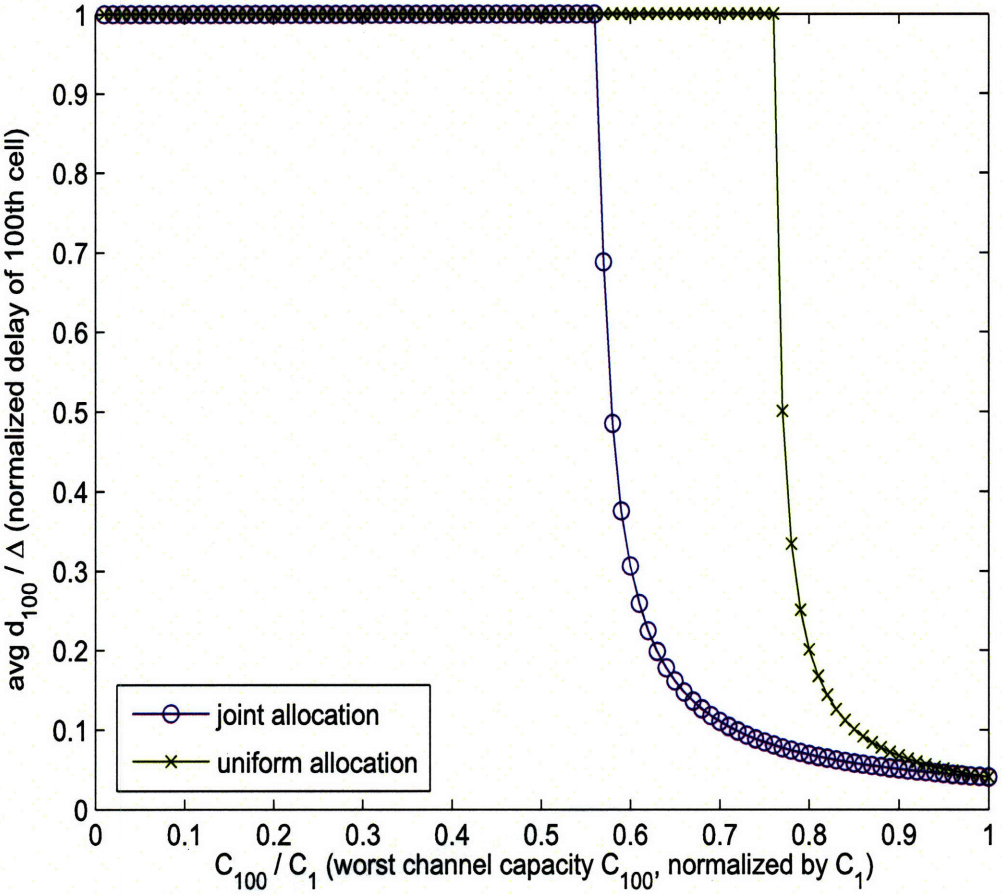

Figure 4-10: Comparing average delays of the $N^{th}(= 100^{th})$ cell normalized by the deadline of joint and uniform beam allocation/congestion control as a function of the worst channel capacity C_{100} (normalized by the fixed capacity of the 1^{st} cell, C_1) for linearly distributed channel capacities

4.4 Impact of Changes of Traffic Demand and Channel Conditions

Here we discuss the impact of the change of external parameters, traffic demand and channel conditions, to the performance of the joint scheme of beam allocation and congestion control, by providing some simple examples.

First, we consider a case where the traffic arrival rates A_i are i.i.d. (independently and identically distributed) with a probability density function (PDF) of $p_A(A_i)$ for

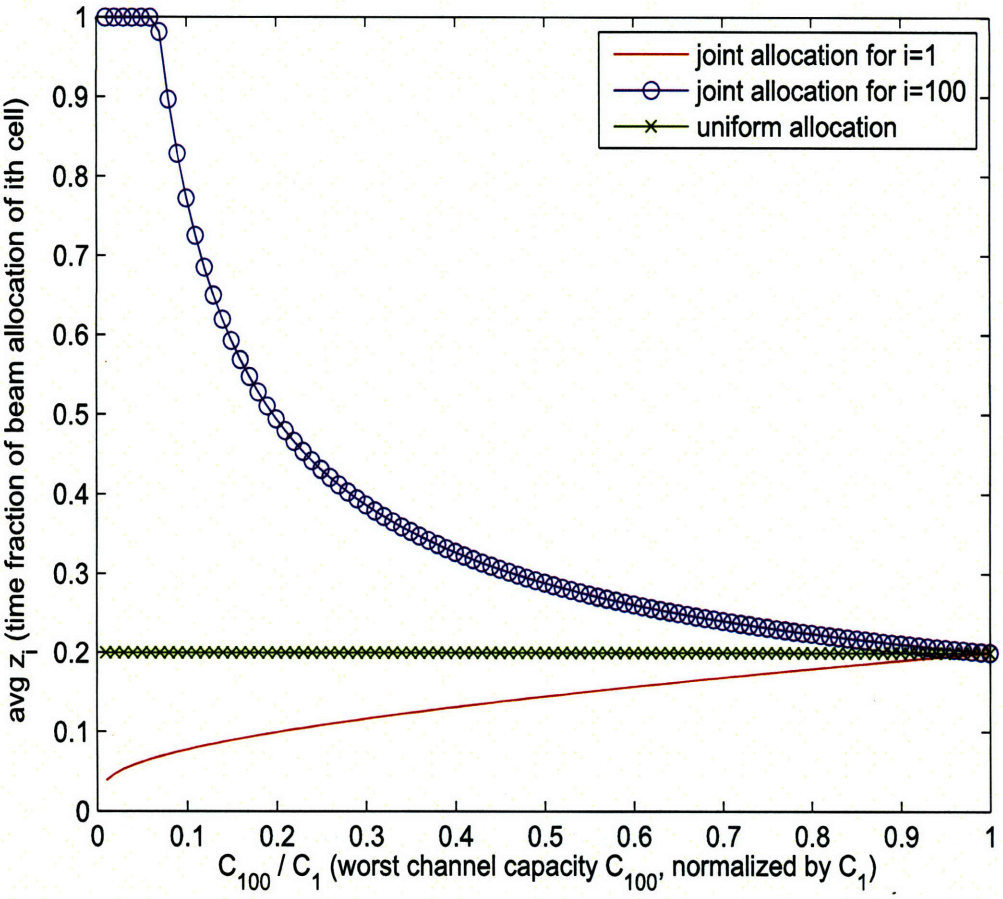

Figure 4-11: Comparing average time fractions of beam allocation of the 1^{st} and $100^{th}(= N^{th})$ cell of joint and uniform beam allocation/congestion control as a function of the worst channel capacity C_{100} (normalized by the fixed capacity of the 1^{st} cell, C_1) for linearly distributed channel capacities

$i = 1, \cdots, N$. Due to the symmetry of the cells, A_N is assumed to be the maximum among A_i's without loss of generality. With the assumption of $C_i \equiv C$ and $\Delta_i \equiv \Delta$ for every i, we define

$$S \equiv \sum_{i=1}^{N-1} A_i, \tag{4.44}$$

which has a PDF of $p_{S|A_N \max}(S)$, a convolution product of $N-1$ PDFs of $p_{A|A_N \max}(A_i)$. Note that every PDF should be conditional to the maximum A_N due to our assumption. We define several constants:

$$X \equiv C\left(1 - \frac{1}{C\Delta}\right) \tag{4.45}$$

and

$$Y \equiv NC\left(\frac{K}{N} - \frac{1}{C\Delta}\right). \tag{4.46}$$

For large N, we can claim that $Y > X$. Then, with A_N maximum, from Eq. (4.25) we have

$$\theta_{joint} = \begin{cases} \frac{X}{A_N} & \text{if } A_N > S \cdot \frac{Y}{Y-X} \text{ and } A_N > Y \\ \frac{Y}{A_N+S} & \text{if } A_N \leq S \cdot \frac{Y}{Y-X} \text{ and } A_N + S > X \\ 1 & \text{if } A_N \leq X \text{ and } A_N + S \leq X \end{cases}, \tag{4.47}$$

which is shown in Fig. 4-12. Note that $\{(A_N, S) | S > (N-1)A_N\}$ is not a feasible region since A_N is assumed to be the maximum amongst A_i's.

Again, this example confirms that θ depends on the traffic distributions of all the cells. With a small number of heavily demanding cells and the corresponding $\theta_{joint} = \frac{X}{A_N}$, the multiple beam antenna loses some efficiency compared to that with more balanced traffic among every cell and the corresponding $\theta_{joint} = \frac{Y}{A_N+S}$. The drawback can be overcome by the use of a phased array antenna, which is studied in Chapter 5.

The expected value of θ_{joint} can be obtained by

$$E[\theta_{joint}] = \sum_{i=1}^{N} E[\theta_{joint} | A_i \max] \cdot \Pr[A_i \max]$$

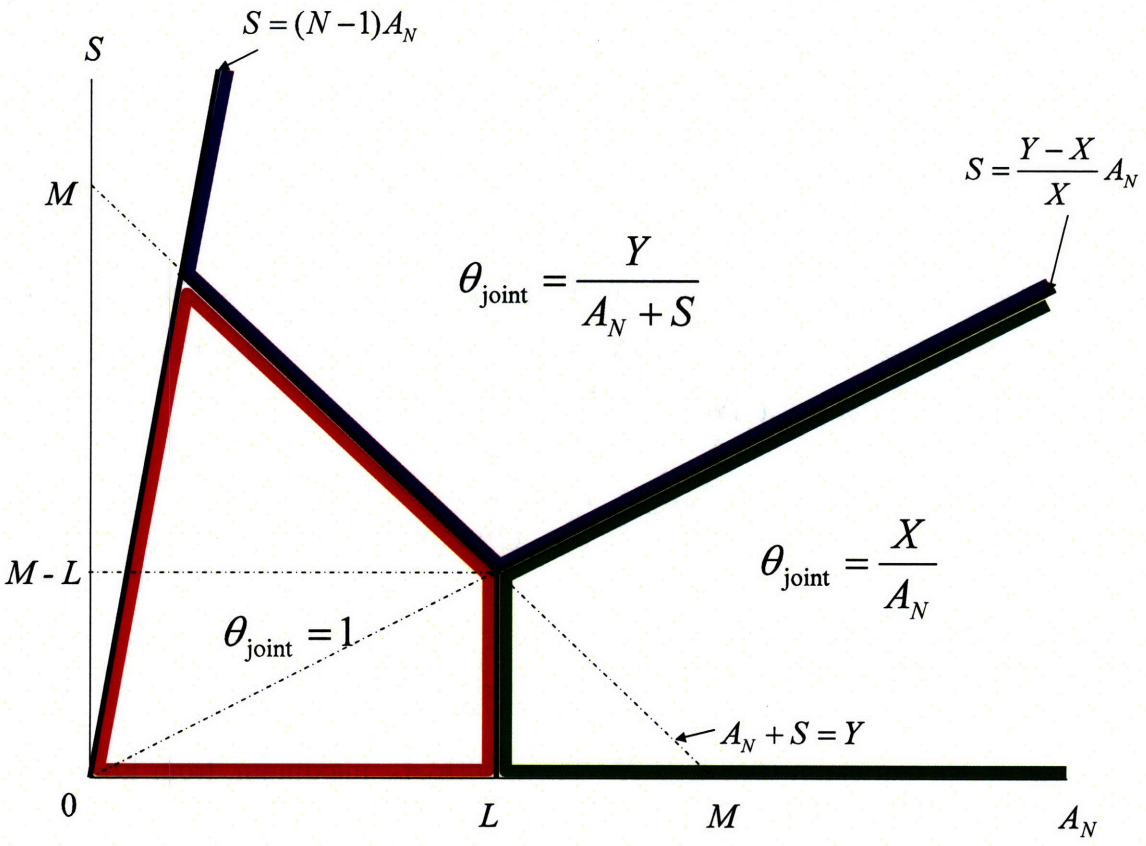

Figure 4-12: The value of θ_{joint} according to traffic distribution of the most dominant user (in terms of A_N) and others (in terms of $S = \sum_{i=a}^{N-1} A_i$)

$$\begin{aligned} &= E[\theta_{joint} | A_N \max] \\ &= \int \theta_{joint} \cdot p_{A_N, S | A_N \max}(A_N, S) dA_N dS, \end{aligned} \tag{4.48}$$

with a conditional PDF $p_{A_N,S|A_N \max}(A_N, S)$. The second equality comes from the symmetry of every cell.

Next, we consider the example with linearly distributed traffic arrival rates

$$A_i = i \cdot \beta \tag{4.49}$$

and channel capacities

$$C_i = i \cdot \phi \tag{4.50}$$

for every cell $i = 1, \cdots, N$. Every cell has the same utilization factor of β/ϕ (before scaled by θ and z_i). With $\Delta_i \equiv \Delta$ for every i, from Eq. (4.25), we have

$$\theta_{joint} = \begin{cases} \frac{1}{\beta} \cdot \left(\phi - \frac{1}{\Delta}\right) & \text{if } \phi < \beta + \frac{1}{\Delta} \text{ and } \phi < \frac{1}{\Delta} \cdot \frac{N - \sum_i i^{-1}}{N-K} \\ \frac{1}{\beta} \cdot \frac{K}{N} \cdot \left(\phi - \frac{1}{K\Delta} \sum_i i^{-1}\right) & \text{if } \phi > \beta \cdot \frac{N}{K} + \frac{1}{K\Delta} \sum_i i^{-1} \text{ and } \phi \geq \frac{1}{\Delta} \cdot \frac{N - \sum_i i^{-1}}{N-K} \\ 1 & \text{if } \phi \geq \beta + \frac{1}{\Delta} \text{ and } \phi \leq \beta \cdot \frac{N}{K} + \frac{1}{K\Delta} \sum_i i^{-1} \end{cases}, \tag{4.51}$$

which is shown in Fig. 4-13. Since we can show

$$\sum_{i=1}^{N} i^{-1} \sim \ln N < K, \tag{4.52}$$

from $\sum_{i=1}^{100} i^{-1} = 5.19$ and $\sum_{i=1}^{1000} i^{-1} = 7.49$, we obtain the order of the following three values:

$$\frac{1}{K\Delta} \sum_{i=1}^{N} i^{-1} < \frac{1}{\Delta} < \frac{1}{\Delta} \cdot \frac{N - \sum_{i=1}^{N} i^{-1}}{N-K}, \tag{4.53}$$

each of which is the intersection point between the ϕ-axis and the boundary line that decides θ_{joint}.

As the channel condition becomes better compared to traffic demand (near the ϕ-axis), the system can admit all the incoming traffic with $\theta_{joint} = 1$. On the other hand, if the channel condition becomes worse compared to traffic demand (near the β-axis), the cell with the worst channel condition, $N = 1$, is a bottleneck for the system and decides θ_{joint} because the additional price $\frac{1}{C_i \Delta_i}$ for $\bar{z}_i$ in Eq. (4.26) is the highest for user 1 with the identical A_i/C_i and Δ_i among all the users. Between two regions, θ_{joint} is decided by considering the whole system. Even in the idealistic example of linearly distributed traffic demand and channel capacities, the changes of the two external parameters result in significant performance difference.

4.5 Summary

To meet average delay constraints and to stabilize the system, we should consider some form of congestion control of incoming traffic and couple it with resource allo-

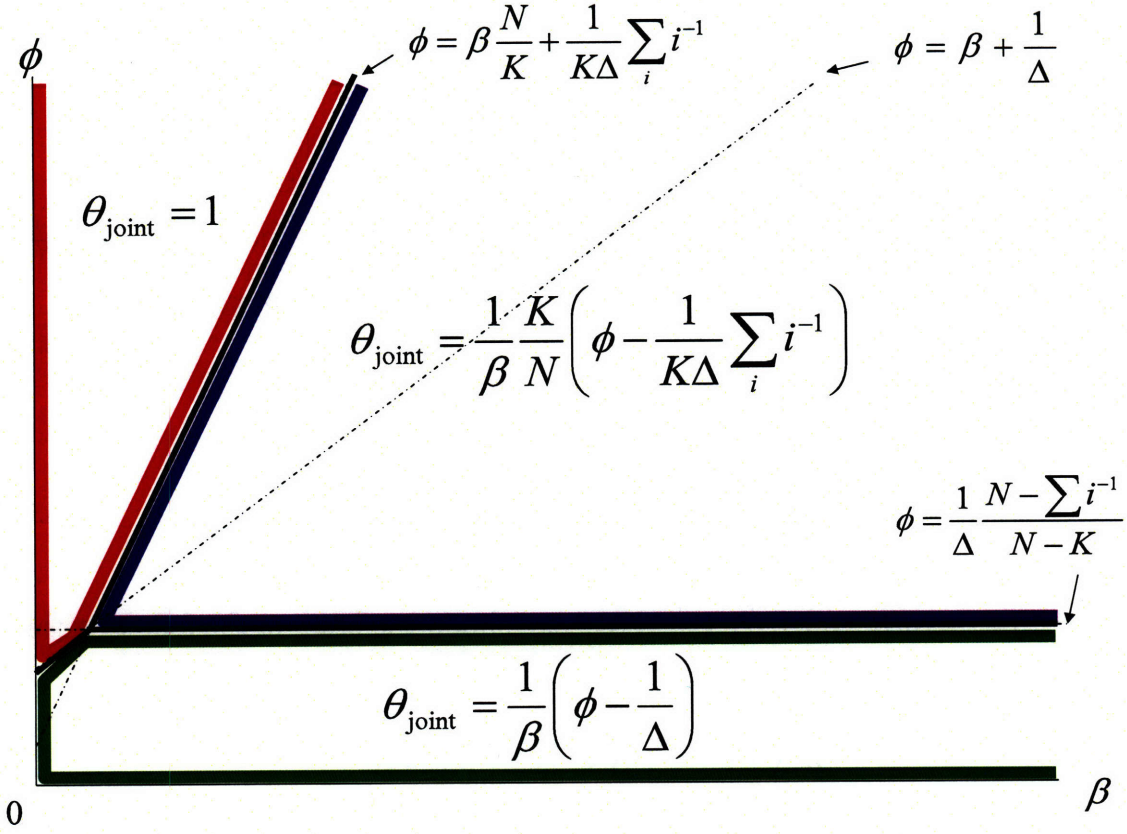

Figure 4-13: The value of θ_{joint} according to traffic distribution (in terms of a parametric slope β of the linear traffic among users) and channel conditions (in terms of a parametric slope ϕ of the linear channel capacities among users)

cation. This is more important for data communications with bursty and unscheduled computer traffic. In this chapter, for the multiple beam antenna, we have found the optimal solution for joint multibeam allocation and congestion control over satellite downlinks based on incoming traffic, link qualities and average delay constraints. We have modeled a maximization problem of throughput by assuming quasi-static channel conditions and very fast beam switching techniques. We have analytically found the optimal congestion control parameter and corresponding beam allocation method. The comparison with uniform allocation has shown that the joint scheme has advantages of more accepted incoming traffic and/or a smaller queueing delay.

In the practical system with finite-length queues, if there is no congestion control mechanism, excessive packet arrival will result in packet loss after the queue becomes

full, and can initiate unnecessary ARQ functions and retransmissions for dropped packets, inducing possibly more congestion. Congestion control can avoid this by regulating the amount of incoming traffic with an acceptable queueing delay.

Chapter 5

Joint Phased Array Antenna Gain Patterning And Scheduling

In Chapter 4, we presented a problem of multiple beam allocation and congestion control, assuming the use of multiple beam antenna with traveling wave tube amplifiers (TWTA). Each multiple beam antenna feed is fed by its own TWTA, which results in a power constraint for each beam. Since it is assumed that TWTAs are driven well into saturation for efficiency with a single carrier, we can fix power at the maximum possible level when a TWTA is operating. The channel condition is quasi-static during the interval of interest, and the channel capacity for each beam is considered to be constant. In this chapter, we consider the use of phased array antenna (Fig. 5-1). A phased array antenna uses solid state power amplifiers (SSPA) and can linearly superimpose signals at array elements by controlling an antenna-patterning matrix. Signal power can be divided among multiple channels up to the total power of the array. The time-varying channel capacity is not full-on or off any more since the way of allocating power is flexible in the phased array antenna. In addition, while the multiple beam antenna has fixed beam size due to the fixed size of feedhorn for each signal, the phased array antenna can have any size and/or shape of beam by feeding many array elements with the same signal. Moreover, the phased array antenna together with transmission scheduling can be cycled much more rapidly ($\ll$

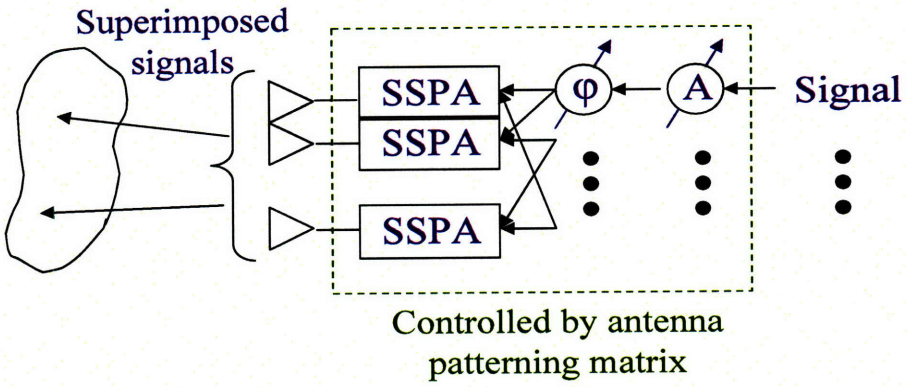

Figure 5-1: A schematic of phased array antenna

msec) than the multiple beam antenna and is advantageous in meeting time deadlines via fast switching of resources. The optimum design of antenna gain patterning and scheduling of the phased array antenna adaptive to traffic distribution and channel conditions can enhance the efficiency of satellite resource allocation.

In Section 5.1, we formulate a resource allocation problem for a phased array antenna system. By considering two extreme cases of (i) widely spread users and (ii) very close-in users, we derive optimum antenna gain patterning in Section 5.2 and beam scheduling in Section 5.3. Specific examples show that the choice of the optimum scheme depends on user distribution and signal-to-noise ratios. In Section 5.4, the performance of the phased array antenna is compared with that of the multiple beam antenna. We develop an efficient algorithm for user selection, antenna gain patterning, power allocation and admission control in Section 5.5. Simulation results are given and compared with the steady-state solution in Section 5.6. We summarize the chapter in Section 5.7.

5.1 Formulation

We assume that there are M users[1] on the Earth coverage area and each user expects to receive a different signal from a satellite with a phased array antenna. On the antenna aperture plane (ξ, η), which is assumed to be continuous over $|\xi| \leq \frac{D}{2}$ and $|\eta| \leq \frac{D}{2}$, the amplitudes and phases of array elements are controlled by a pattern-forming complex, to synthesize K ($\ll M$) active downlink signals (Fig. 5-2), whose number is limited by the number of onboard modulators. Denote the field distribution at the aperture as $V_i(\xi, \eta)$ for the i^{th} user. The field distributions for all users are linearly superimposed, and we have the total field distribution of the antenna element at (ξ, η), given as

$$V_{sum}(\xi, \eta) = \sum_{i=1}^{M} V_i(\xi, \eta). \tag{5.1}$$

The aperture power density transmitted at (ξ, η) is $|V_{sum}(\xi, \eta)|^2$. On the antenna plane, each element has the maximum power density constraint of

$$|V_{sum}(\xi, \eta)|^2 \leq \rho_0, \tag{5.2}$$

for a constant ρ_0.

For a transmit antenna of width D, wavelength λ, and altitude of the satellite L, the minimum beamwidth of the mainlobe illuminated by one diffraction-limited beam is $\frac{\lambda L}{D}$. From the Fraunhofer diffraction approximation [26] in the case of far-field transmission with $\frac{D^2}{\lambda} \ll L$, the received signal $U_{sum}(x, y)$ on the Earth surface (x, y) is given by the two-dimensional Fourier transform of the field distribution $V_{sum}(\xi, \eta)$, written as

$$U_{sum}(x, y) = \frac{e^{j\frac{2\pi L}{\lambda}} e^{j\frac{\pi}{\lambda L}(x^2+y^2)}}{j\lambda L} \int\int V_{sum}(\xi, \eta) e^{-j\frac{2\pi}{\lambda L}(x\xi+y\eta)} d\xi d\eta, \tag{5.3}$$

where a simple path loss is only considered.

[1] In this chapter, we focus on the number of users M instead of that of cells N with the phased array antenna. This is discussed in more detail in Section 5.2.

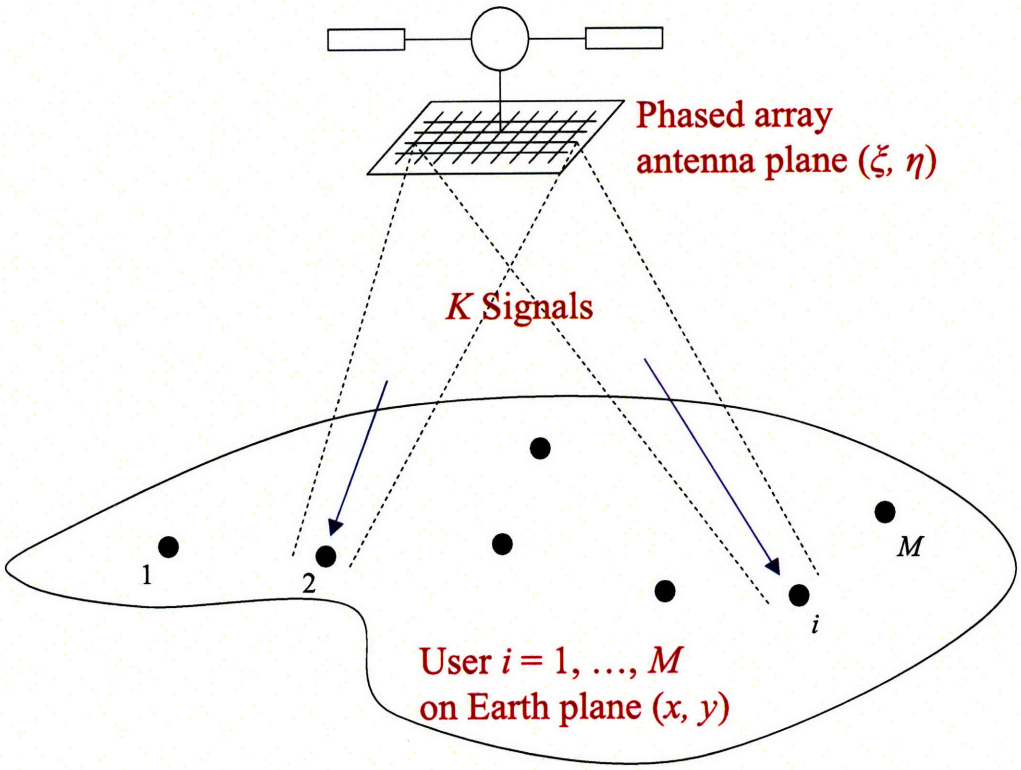

Figure 5-2: A phased array antenna satellite, generating K active signals and serving M users on the Earth

Since $V_{sum}(\xi, \eta)$ is the linear superposition of $V_i(\xi, \eta)$ that is Fourier-transformed to $U_i(x, y)$ for each i, we have the received signal of

$$U_{sum}(x, y) = \sum_{i=1}^{M} U_i(x, y). \tag{5.4}$$

Every received signal waveform at (x, y) is assumed to have the identical propagation delay because the signal comes from the same satellite that is located very far. Every user is also assumed to equip with the same unit size of receiver antenna, so its impact is ignored.[2]

In the multiuser communications theory [47], the performance of the system depends on the type of receiver for detecting the desired signal. The optimum maximum-likelihood (ML) receiver has exponential computation complexity in terms of the

[2]Received signal power is proportional to the receiver antenna size, as in Eq. (3.1) in Chapter 3.

number of users M. Instead, simple linear receivers [52, 53] can be used, such as a minimum mean square error (MMSE) detector, which maximizes the output signal-to-interference-and-noise ratio (SINR) among the family of linear receivers. Moreover, if signals are assumed to be jointly Gaussian, the linear MMSE receiver becomes the optimum ML receiver. By using an MMSE receiver, we have the signal-to-interference-and-noise-ratio (SINR) of

$$SINR_i = \frac{P_i(x_i, y_i)}{WN_0 + \sum_{k \neq i} P_k(x_i, y_i)}, \tag{5.5}$$

where $P_i(x_i, y_i) = |U_i(x_i, y_i)|^2$ is the received power of the i^{th} signal at (x_i, y_i), W is the bandwidth used, and N_0 is the white noise spectral density. Note that all the other signals except the i^{th} are treated as interference. By the use of error correction codes, we assume that we can achieve close to the Shannon capacity with the given SINR. We incorporate signal power attenuation due to the weather effects that only change quasi-statically, α_i^2 (≤ 1) over the link to the i^{th} user. The capacity achievable for the i^{th} user at (x_i, y_i) is

$$C_i = W \log \left(1 + \frac{\alpha_i^2 P_i(x_i, y_i)}{WN_0 + \sum_{k \neq i} \alpha_i^2 P_k(x_i, y_i)} \right). \tag{5.6}$$

Only K users out of M can be served at one time (excluding the case of broadcasting same signals to different users). We have two control variables: (i) signal assignment $z_i(t)$, which indicates whether the i^{th} user receives the signal at time t, and (ii) aperture field distribution $V_i(\xi, \eta, t)$, which is translated to power allocation $P_i(x, y, t)$. Different from the multiple beam antenna with binary (on-off) power allocation to each beam (by driving TWTAs up to saturation for best efficiency), the phased array antenna has multi-valued power allocation, and thus the channel capacity in Eq. (5.6) is also multi-valued, which makes the problem complicated. The time-average queueing delay $\bar{d}_i$ of the packets for the i^{th} user is the function of the channel capacity $z_i \cdot C_i$ (equal to 0 if the user has no signal assigned with $z_i = 0$) and the average rate of incoming traffic A_i.

If our metric is to maximize the congestion control parameter θ of incoming traffic (and thus the throughput) as in Chapter 4, the problem is given as

$$\text{maximize } \theta \tag{5.7}$$

$$\text{subject to } \quad 0 \leq \theta \leq 1 \tag{5.8}$$

$$\bar{d}_i(\theta A_i, z_i(t)C_i(t)) \leq \Delta_i \tag{5.9}$$

$$\sum_{i=1}^{M} z_i(t) \leq K \qquad (z_i(t) = 0 \text{ or } 1 \text{ for } t \text{ and } i) \tag{5.10}$$

$$C_i(t) = W \log\left(1 + \frac{\alpha_i^2 P_i(x_i, y_i, t)}{WN_0 + \sum_{k \neq i} \alpha_i^2 P_k(x_i, y_i, t)}\right) \tag{5.11}$$

$$P_i(x, y, t) = |U_i(x, y, t)|^2 \tag{5.12}$$

$$U_i(x, y, t) = \mathcal{F}[V_i(\xi, \eta, t); L] \tag{5.13}$$

$$V_{sum}(\xi, \eta, t) = \sum_{i=1}^{M} V_i(\xi, \eta, t) \tag{5.14}$$

$$|V_{sum}(\xi, \eta, t)|^2 \leq \rho_0 \quad \text{if } |\xi| \leq \tfrac{D}{2} \text{ and } |\eta| \leq \tfrac{D}{2} \tag{5.15}$$

$$\text{and} \quad V_{sum}(\xi, \eta, t) = 0 \quad \text{if } |\xi| > \tfrac{D}{2} \text{ or } |\eta| > \tfrac{D}{2}. \tag{5.16}$$

Average delay and maximum K independent signal constraints are shown in (5.9) and (5.10) respectively. Δ_i (> 0) is a given delay deadline for the i^{th} user. $\mathcal{F}[\cdot; L]$ in (5.13) is a two-dimensional far-field Fourier transform with distance L between the field distribution at the antenna aperture and the received signal as in (5.3). Multiple signals are linearly superimposed in (5.14). The power density constraint at every antenna element is shown in (5.15) and (5.16).

Solving this optimization problem gives the joint optimum solution of

- how to schedule downlink transmissions and select K active users in terms of $P_i(t)$ and $z_i(t)$, and

- how to pattern the antenna aperture distribution in terms of $V_i(\xi, \eta)$, in order

to reduce interference between active users and maximize the SINR.

We first try to understand feasible solutions to separate sub-problems and then give a joint solution. In Section 5.2 we describe the optimum antenna gain patterning of (ξ, η) with the scheduling part suppressed, i.e., assuming that K users with $z_i(t) = 1$ are given. We focus on reducing interference to maximize the SINR with given transmit power. If interference is too severe between close-in users, sequential service should be deployed, which is in fact a scheduling problem. With the knowledge of antenna gain patterning, we obtain the optimum scheduling policy in Section 5.3: which K users are selected each time and how much transmit power is given to each user. Then, we will show that these two decisions are eventually combined and made jointly depending on user location/demand and channel conditions.

5.2 Antenna Gain Patterning

We begin the section with the special case of a single active signal, i.e., $K = 1$, and then move to the general case of multiple signal transmission. We consider two ways of mitigating interference between multiple signals: antenna gain patterning for interference suppression and scheduling spatially orthogonal patterns for negligible interference. In addition to conventional multiplexing methods of time, frequency and code division multiplexing (TDM, FDM and CDM), we add space division multiplexing (SDM) even for the very close users whose mainlobes can overlap, by generating interference-suppressed signal patterns appropriately from the satellite transmitter with a phased array antenna. To support SDM, the coverage area of a multiple beam satellite is covered by a number of spotbeams, which are considered as cells in a cellular system. This is the model that we have considered in Chapter 3 and 4. An SDM scheme using multiple spotbeams assigns the same frequency to non-adjacent spotbeams, so that one can reuse frequency and increase system capacity. In this section, we discard the concept of the cellular system with fixed-size cells illuminated by fixed size beams, but focus on individual user locations since the phased array

antenna can provide flexible beam size and shape (within the limitation of diffraction theory). We will show that, for some distance range between active users, SDM with interference suppression can outperform the orthogonal schemes of TDM and FDM.

A conventional CDM satellite downlink transmission forms simultaneous signals in a single type of modulation and performs power control to guarantee the same level of received signal power, mainly compensating for the signal attenuation due to channel degradation. An advanced system will deploy adaptive modulation that changes the symbol size according to channel conditions and traffic load, in order to save precious onboard power and/or to optimize data rates [8, 9]. If two close users have very different symbol size and thus, different received power levels, the weaker signal can be overshadowed by the interference from the stronger signal since signature codes cannot be orthogonal to each other all the time. A complicated power balance control may solve this problem, but not be applicable to the satellite data transmission with long propagation delay and diverse user demand. Antenna gain patterning with interference suppression is the only viable solution in this case, and thus, we consider the possibly worst-case scenario in the advanced adaptive modulation system (e.g., military satellites).

5.2.1 Single Beam Transmission

We assume that the satellite has only single active transmission. All the antenna elements can be used for the single signal and we do not need to consider interference. The maximum power to the desired user results in the optimum performance such as the maximum capacity and the smallest delay. To maximize the received power at the desired location (x_0, y_0), we maximize the signal intensity of

$$|U(x_0, y_0)| = \frac{1}{\lambda L} \left| \int \int V(\xi, \eta) e^{-j\frac{2\pi}{\lambda L}(x_0\xi + y_0\eta)} d\xi d\eta \right| \tag{5.17}$$

subject to the transmit power constraint $|V(\xi,\eta)|^2 \leq \rho_0$. We express the aperture distribution as

$$V(\xi,\eta) = G(\xi,\eta)e^{j\omega(\xi,\eta)} \tag{5.18}$$

where $G(\xi,\eta)$ and $\omega(\xi,\eta)$ are the amplitude and phase component respectively on the antenna aperture. Then, we have

$$\begin{aligned} |U(x_0,y_0)| &= \frac{1}{\lambda L}\left|\int\int V(\xi,\eta)e^{-j\frac{2\pi}{\lambda L}(x_0\xi+y_0\eta)}d\xi d\eta\right| \\ &\leq \frac{1}{\lambda L}\int\int\left|V(\xi,\eta)e^{-j\frac{2\pi}{\lambda L}(x_0\xi+y_0\eta)}\right|d\xi d\eta \\ &= \frac{1}{\lambda L}\int\int G(\xi,\eta)d\xi d\eta, \end{aligned} \tag{5.19}$$

where the equality holds when

$$\omega(\xi,\eta) = \frac{2\pi}{\lambda L}(x_0\xi + y_0\eta). \tag{5.20}$$

Since the total transmit power is fixed, i.e.,

$$\int\int |V(\xi,\eta)|^2 d\xi d\eta = \int\int G^2(\xi,\eta)d\xi d\eta = P_{total}^T, \tag{5.21}$$

we can maximize $\frac{1}{\lambda L}\int\int G(\xi,\eta)d\xi d\eta$ by solving the Lagrangian functional of

$$J(G) = \frac{1}{\lambda L}\int\int G(\xi,\eta)d\xi d\eta - \Lambda\left(\int\int G^2(\xi,\eta) - P_{total}^T\right)d\xi d\eta \quad \text{with } G(\xi,\eta) \geq 0, \tag{5.22}$$

where Λ (> 0) is a Lagrange multiplier. Differentiating the Lagrangian functional with respect to G gives

$$G(\xi,\eta) = \frac{1}{2\Lambda\lambda L}, \tag{5.23}$$

which is a constant decided by the maximum power density constraint (5.15), such that

$$G(\xi,\eta) = \sqrt{\rho_0} \quad \text{for every element.} \tag{5.24}$$

Thus, the optimum field distribution is given as

$$V(\xi,\eta) = \sqrt{\rho_0}\, e^{j\frac{2\pi}{\lambda L}(x_0\xi + y_0\eta)} \quad \text{for every element.} \tag{5.25}$$

The optimum distribution has the constant amplitude $\sqrt{\rho_0}$ over the entire aperture and the linear phase component of shifting the maximum power to the desired location by using a directional vector in

$$\omega(\xi,\eta) = \frac{2\pi}{\lambda L}(x_0, y_0)\cdot(\xi,\eta). \tag{5.26}$$

The received power from $V(\xi,\eta)$ in Eq. (5.25) gives

$$\begin{aligned} P(x,y) &= |U(x,y)|^2 \\ &= \left[\frac{1}{\lambda L}\left|\int\int V(\xi,\eta)e^{-j\frac{2\pi}{\lambda L}(x\xi+y\eta)}d\xi d\eta\right|\right]^2 \\ &= \left[\frac{1}{\lambda L}\left|\int_{-D/2}^{D/2}\int_{-D/2}^{D/2}\sqrt{\rho_0}e^{-j\frac{2\pi}{\lambda L}\{(x-x_0)\xi+(y-y_0)\eta\}}d\xi d\eta\right|\right]^2 \\ &= \frac{\rho_0 D^4}{\lambda^2 L^2}\mathrm{sinc}^2\left[\frac{D(x-x_0)}{\lambda L}\right]\mathrm{sinc}^2\left[\frac{D(y-y_0)}{\lambda L}\right], \end{aligned} \tag{5.27}$$

where a sinc function is defined by

$$\mathrm{sinc}\ x = \begin{cases} 1 & \text{if } x = 0 \\ \frac{\sin \pi x}{\pi x} & \text{if } x \neq 0. \end{cases} \tag{5.28}$$

The two-dimensional sinc function with x and y independent of each other is due to the use of square (D by D in this case) transmit antenna. If a circular antenna is used, the solution is a form of Bessel function [2]. The two-dimensional sinc function of Eq. (5.27) has the narrowest beamwidth of the main lobe, which is $\frac{\lambda L}{D}$, amongst all the beam patterns illuminated by the limited size of square aperture [33], and gives the highest received power in the wanted location for a given transmit power. Suppose that the user at the desired location has a receiver of δ-by-δ square. Since the receiver

is much smaller than the satellite beam, the received power can be approximated by

$$P(x_0, y_0) \cdot \delta^2 = \frac{\rho_0 D^4 \delta^2}{\lambda^2 L^2} = \frac{D^2 \delta^2}{\lambda^2 L^2} P_{total}^T, \tag{5.29}$$

where the available total transmit power P_{total}^T is given as

$$P_{total}^T = \int\int |V(\xi, \eta)|^2 d\xi d\eta = \rho_0 D^2. \tag{5.30}$$

The optimum aperture distribution has the ratio of $\frac{D^2\delta^2}{\lambda^2 L^2}$ between the transmitted and received power. This is the same as in Chapter 3 where it is ideally assumed that the beam has a uniform power density inside the diffraction limited distance $\frac{\lambda L}{D}$ and no side lobe outside. It is reminded that in Chapter 3 we ignored interference and only focused on power allocation, implicitly assuming the optimum gain patterning derived in this section for multiple beams with negligible interference.

5.2.2 Multiple Beam Transmission for Sparse Users

We now consider the case of transmitting multiple K (> 1) independent active signals. The signal on each antenna element is multiplied by a time-varying waveform $v_i(t)$ that is a product of a unit-power signature waveform and binary information data, i.e.,

$$V_i(\xi, \eta, t) = v_i(t) V_i(\xi, \eta). \tag{5.31}$$

We assume that every $v_i(t)$ is independent of each other and its time average is equal to zero ($\bar{v_i}(t) = 0$ but $\overline{|v_i(t)|^2} = 1$ where $\bar{x}$ represents a time average of x). The change rate of temperature in the SSPA due to heating is of the order of milliseconds while the signal change speed of $v_i(t)$ can be of the order of microseconds or smaller. Thus, the heat accumulation from the linear term of V_i averages out to be negligible before it changes the temperature of the SSPA. That is, in the power constraint of

$$\left|\sum V_i(\xi, \eta, t)\right|^2 = \sum |V_i(\xi, \eta, t)|^2 + \sum_{i \neq k} V_i(\xi, \eta, t) V_k^*(\xi, \eta, t) \leq \rho_0, \tag{5.32}$$

we can ignore heat accumulation of

$$\sum_{i \neq k} V_i(\xi, \eta, t) V_k^*(\xi, \eta, t) = \sum_{i \neq k} v_i(t) v_k^*(t) V_i(\xi, \eta) V_k^*(\xi, \eta) \tag{5.33}$$

from

$$\overline{v_i(t) v_k^*(t)} = 0, \tag{5.34}$$

where x^* represents a complex conjugate of x. Then, we only focus on

$$\sum |V_i(\xi, \eta, t)|^2 \leq \rho_0. \tag{5.35}$$

This approximation suppresses the cross-products of different signals and thus decouples signals in amplitude/phase adjustment at each element. In practice, compared to the TWTA, the SSPA usually has a wider range of linearity before a sharp cut-off, so that linearity can be assured to allow superposition of signals at each element.

Multiple spotbeams can give a higher total throughput than a single beam (by the concavity of the capacity function with respect to power as discussed in Section 3.1) at the expense of possible interference from other active signal patterns. Interference is a monotonically decreasing function of distance between active users for inside the mainlobe and assumed to be negligible outside. That is, we mainly deal with the interference due to the overlapping mainlobes for extremely close users. Though sidelobe interference may not be insignificant, especially from the first sidelobes of adjacent beams much stronger than the beam of interest, the resulting degradation is smaller than that from the mainlobe (e.g., in the sinc function, the power difference is 13.5 dB between the main lobe and the first sidelobe and 17.9 dB between the main lobe and the second sidelobe), and easier to overcome at the smaller cost (e.g., locating nulls for interference without reducing the desired signal power much, or using error correction codes that can restore signals of up to 1 dB SINR).

If active users are located far enough with very small interference over the satellite coverage area $A_{coverage}$ that satisfies $A_{coverage} \gg K \cdot \left(\frac{\lambda L}{D}\right)^2$, we can locate multiple

narrowest spotbeams farther than the smallest beamwidth (of the mainlobe) $\frac{\lambda L}{D}$. This provides the maximum SINR and thus, the maximum throughput because the narrowest spotbeam in the form of (5.25) gives the maximum received power out of the given transmit power for each non-interfering signal by the same argument as in the single beam case.

Here, the optimality of the narrowest spotbeams can be shown even for a general convex utility function $f(C; A)$ in terms of vectors of capacities C and arrival rates A.[3] In this case, the optimum antenna gain patterning problem (with the scheduling problem and the maximum K independent signal constraint suppressed) is restated as

$$\text{maximize } f(C; A) \tag{5.37}$$

$$\text{subject to } \sum_i V_i(\xi, \eta, t) V_i^*(\xi, \eta, t) \leq \rho_0 \quad \text{if } |\xi| \leq \tfrac{D}{2} \text{ and } |\eta| \leq \tfrac{D}{2}. \tag{5.38}$$

The Lagrangian function with a Lagrangian multiplier $\mu(\xi, \eta, t)$ (real nonnegative) is given as

$$J[V_i^*(\xi, \eta, t)] = f - \int\int \mu(\xi, \eta, t) \left[\sum_i V_i(\xi, \eta, t) V_i^*(\xi, \eta, t) - \rho_0 \right] d\xi d\eta, \tag{5.39}$$

where we consider $V_i^*(\xi, \eta, t)$, which is a complex conjugate of $V_i(\xi, \eta, t)$, as a variable while $V_i(\xi, \eta, t)$ as a constant, following the discussion on complex gradients in Van Trees's textbook [54] (pp. 1402 - 1404), which is restated in the chapter appendix, Section 5.8. Differentiating J with respect to $V_i^*(\xi, \eta, t)$ gives

$$\begin{aligned} \frac{\partial J}{\partial V_i^*} &= \frac{\partial f}{\partial C_i} \cdot \frac{\partial C_i}{\partial P_i} \cdot \frac{\partial P_i}{\partial V_i^*} - \mu(\xi, \eta, t) V_i(\xi, \eta, t) \\ &= s_i(t) e^{j \frac{2\pi}{\lambda L}(x_i \xi + y_i \eta)} - \mu(\xi, \eta, t) V_i(\xi, \eta, t) = 0, \end{aligned} \tag{5.40}$$

[3] In our optimization problem of (5.7), the utility function f is given as a system throughput with a delay penalty,

$$f = \theta - \sum_i \kappa_i \cdot (\bar{d}_i - \Delta_i) \tag{5.36}$$

where a Lagrangian multiplier κ_i can be positive only if $\bar{d}_i \geq \Delta_i$ and is equal to zero otherwise.

where

$$s_i(t) \equiv \frac{\partial f}{\partial C_i} \cdot \frac{\alpha_i^2/N_0}{1+\frac{\alpha_i^2 P_i(x_i,y_i,t)}{WN_0}} \cdot \frac{1}{\lambda L} U_i(x_i,y_i,t) \tag{5.41}$$

represents the desired component of the i^{th} signal and is independent of (ξ,η). In (5.40) we use the followings with the unit size receiver antenna ($\delta = 1$):

$$P_i(t) = U_i(x_i,y_i,t)U_i^*(x_i,y_i,t) = U_i(x_i,y_i,t)\frac{1}{\lambda L}\int\int V_i^*(\xi,\eta,t)e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)}d\xi d\eta \tag{5.42}$$

and

$$\frac{\partial P_i}{\partial V_i^*} = \frac{1}{\lambda L}U_i(x_i,y_i,t)e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)}. \tag{5.43}$$

Combining (5.38) and (5.40), and assuming the equality in the power constraint (5.38) at every element for maximum efficiency, we have

$$\sum |V_i(\xi,\eta,t)|^2 = \frac{1}{\mu^2(\xi,\eta,t)}\sum |s_i(t)|^2 = \rho_0. \tag{5.44}$$

Since $|s_i(t)|^2$ is independent of (ξ,η), so is $\mu(\xi,\eta,t) = \sqrt{\frac{\sum |s_i(t)|^2}{\rho_0}}$ $(\equiv \mu(t))$. Every element has the same scaling factor of $\frac{1}{\mu(t)}$ and we have

$$V_i(\xi,\eta,t) = \frac{s_i(t)}{\sqrt{\sum |s_i(t)|^2}} \cdot \sqrt{\rho_0} \cdot e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)}. \tag{5.45}$$

Note that s_i may be a complex variable due to $U_i(x_i,y_i,t)$ in (5.41) and can have a phase term independent of (ξ,η), which does not change the result and thus, is ignored. We treat s_i as a real variable representing the amplitude component (with scaling factor). Each active signal pattern has a uniform amplitude of

$$G_i(t) = \frac{s_i(t)}{\sqrt{\sum |s_i(t)|^2}}\sqrt{\rho_0} \tag{5.46}$$

over the entire aperture and a phase adjusted to point the beam at the desired direction of (x_i,y_i) to add up all the terms in phase. Every active signal is distributed

over all antenna elements because the wider the transmit antenna is, the narrower the mainlobe can be. Multiple signals are linearly superimposed as

$$V_{sum}(\xi,\eta,t) = \sum_{i\in\Omega(t)} \frac{s_i(t)}{\sqrt{\sum_{i\in\Omega(t)}|s_i(t)|^2}} \cdot \sqrt{\rho_0} \cdot e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)} \tag{5.47}$$

for $|\xi| \leq \frac{D}{2}$ and $|\eta| \leq \frac{D}{2}$. $\Omega(t)$ is the set of active signals at time t.

Now, we revisit the results in Chapter 3 on the phased array antenna and show that they are identical to the results derived here. From

$$U_i(x_i,y_i,t) = \frac{1}{\lambda L}\int\int V_i(\xi,\eta,t)e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)}d\xi d\eta = \frac{s_i(t)}{\mu(t)} \cdot \frac{D^2}{\lambda L}, \tag{5.48}$$

we have

$$\frac{1}{\mu(t)} \cdot \left(\frac{D}{\lambda L}\right)^2 \cdot \frac{\partial f}{\partial C_i^{avg}} \frac{\alpha_i^2/N_0}{1+\frac{\alpha_i^2 P_i(x_i,y_i,t)}{WN_0}} = 1. \tag{5.49}$$

This is the same result as in Chapter 3, where $f(C_i) = -\sum(F_i - C_i)^2$ with traffic demand F_i of the i^{th} user. In addition, from

$$P_i(t) \equiv P_i(x_i,y_i,t) = |V_i|^2 \cdot \frac{D^4}{\lambda^2 L^2} \quad \text{and} \quad \sum|V_i|^2 \leq \rho_0, \tag{5.50}$$

we can derive

$$\sum P_i(t) \leq \frac{\rho_0 D^4}{\lambda^2 L^2} \equiv P_{total}, \tag{5.51}$$

which is the same power constraint as we used in Chapter 2 for the phased array antenna.

5.2.3 Multiple Beam Transmission for Close-in Users

Let us assume that multiple users closer than one beamwidth $\frac{\lambda L}{D}$ from each other should be scheduled for services at the same timeslot. This causes significant co-channel interference between close-in users and some form of mitigating interference is necessary for efficient and reliable communications. We will show that the optimum

pattern of each signal depends on the distances between users and their signal-to-noise ratios (SNR), and can be one of three possibilities: (i) complete cancellation of interference, (ii) optimum suppression of interference, and (iii) the sequential service of close-in users.

Interference is formulated in the capacity function as

$$C_i = W \log\left(1 + \frac{\alpha_i^2 P_i(x_i, y_i, t)}{WN_0 + \sum_{k \neq i} \alpha_i^2 P_k(x_i, y_i, t)}\right), \tag{5.52}$$

which is the same as in (5.11). Since the focus is now on the interference intensity $U_i(x_k, y_k, t)$ that gives $P_i(x_k, y_k, t) = |U_i(x_k, y_k, t)|^2$, we add a term of

$$\begin{aligned} 2\lambda L \cdot \Re\left[\sum_i \sum_{k \neq i} \gamma_{ik}(t) U_i^*(x_k, y_k, t)\right] = \\ \lambda L \sum_i \sum_{k \neq i} \gamma_{ik}(t) U_i^*(x_k, y_k, t) + \lambda L \sum_i \sum_{k \neq i} \gamma_{ik}^*(t) U_i(x_k, y_k, t) \end{aligned} \tag{5.53}$$

to the Lagrangian function of (5.39) by using complex Lagrangian multipliers $\gamma_{ik}(t)$. $\Re[x]$ represents a real part of x. Only the real part is taken because all the terms in the Lagrangian function of (5.39) are real. A constant $2\lambda L$ is multiplied for simplicity of calculation.

We will have two cases: whether interference $U_i(x_k, y_k, t)$ is equal to zero or not. If $U_i(x_k, y_k, t) \neq 0$, we force $\gamma_{ik}(t) = 0$ and the value of the Lagrangian function does not change. We solve the same maximization problem as (5.37) and have

$$\frac{\partial J}{\partial V_i^*} = \frac{\partial f}{\partial C_i}\frac{\partial C_i}{\partial V_i^*} + \sum_{k \neq i}\left[\frac{\partial f}{\partial C_k}\frac{\partial C_k}{\partial V_i^*} + \gamma_{ik}(t) e^{j\frac{2\pi}{\lambda L}(x_k \xi + y_k \eta)}\right] - \mu(\xi, \eta, t) V_i(\xi, \eta, t) = 0, \tag{5.54}$$

where we have

$$\begin{aligned} \frac{\partial C_i}{\partial V_i^*} &= \frac{W}{1 + \frac{\alpha_i^2 P_i(x_i, y_i, t)}{WN_0 + \sum_{h \neq i} \alpha_i^2 P_h(x_i, y_i, t)}} \cdot \frac{1}{WN_0 + \sum_{h \neq i} \alpha_i^2 P_h(x_i, y_i, t)} \\ &\quad \cdot \alpha_i^2 \frac{1}{\lambda L} U_i(x_i, y_i, t) e^{j\frac{2\pi}{\lambda L}(x_i \xi + y_i \eta)} \end{aligned} \tag{5.55}$$

and

$$\frac{\partial C_k}{\partial V_i^*} = -\frac{W}{1+\frac{\alpha_k^2 P_k(x_k,y_k,t)}{WN_0+\sum_{h\neq k}\alpha_k^2 P_h(x_k,y_k,t)}} \cdot \frac{\alpha_k^2 P_k(x_k,y_k,t)}{[WN_0+\sum_{h\neq k}\alpha_k^2 P_h(x_k,y_k,t)]^2} \cdot \alpha_k^2 \frac{1}{\lambda L} U_i(x_k,y_k,t) e^{j\frac{2\pi}{\lambda L}(x_k\xi+y_k\eta)}. \quad (5.56)$$

First, we consider the case of complete cancellation with $U_i(x_k, y_k, t) = 0$ for active users i and k ($k \neq i$). Complete cancellation of interference leads to $\frac{\partial C_k}{\partial V_i^*} = 0$, and from (5.54) we have

$$\mu(\xi,\eta,t)V_i(\xi,\eta,t) = s_i(t)e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)} + \sum_{k\neq i}\gamma_{ik}(t)e^{j\frac{2\pi}{\lambda L}(x_k\xi+y_k\eta)}, \quad (5.57)$$

where $s_i(t)$ is the same as in (5.41). Unlike the case of sparse users in (5.47), $V_i(\xi, \eta, t)$ has more than one terms with a different phase for each. The first term of $s_i e^{j\frac{2\pi}{\lambda L}(x_i\xi+y_i\eta)}$ maximizes the signal power at the desired location as in (5.47). The remaining terms with adjustable γ_{ik} cancel interference toward (x_k, y_k) caused by the first term (of user i). Assuming $\mu(\xi, \eta, t) = \mu(t)$ to be constant over every (ξ, η) for simplicity of calculation and only two active users of i and k within one beamwidth, we apply the zero interference constraint of

$$U_i(x_k,y_k,t) = \frac{1}{\lambda L}\int\int V_i(\xi,\eta,t)e^{-j\frac{2\pi}{\lambda L}(x_k\xi+y_k\eta)}d\xi d\eta = 0 \quad (5.58)$$

and obtain

$$\begin{aligned}\gamma_{ik}(t) &\approx -\frac{1}{D^2}\cdot s_i(t)\int_{-D/2}^{D/2}\int_{-D/2}^{D/2} e^{j\frac{2\pi}{\lambda L}\{(x_i-x_k)\xi+(y_i-y_k)\eta\}}d\xi d\eta \\ &= -s_i(t)\cdot \mathrm{sinc}\left[\frac{D(x_i-x_k)}{\lambda L}\right]\mathrm{sinc}\left[\frac{D(y_i-y_k)}{\lambda L}\right] \quad (5.59)\end{aligned}$$

and

$$\mu^2 = \frac{|s_i|^2+|s_k|^2}{\rho_0}\left(1-\mathrm{sinc}^2\left[\frac{D(x_i-x_k)}{\lambda L}\right]\mathrm{sinc}^2\left[\frac{D(y_i-y_k)}{\lambda L}\right]\right). \quad (5.60)$$

However, these interference cancellation terms reduce the desired signal power $P_i(x_i, y_i, t)$, given as

$$\begin{aligned} P_i(x_i, y_i, t) &= \frac{|s_i(t)|^2}{\mu^2(t)} \cdot \frac{D^4}{\lambda^2 L^2} \cdot \left(1 - \text{sinc}^2\left[\frac{D(x_i - x_k)}{\lambda L}\right] \text{sinc}^2\left[\frac{D(y_i - y_k)}{\lambda L}\right]\right)^2 \\ &= P_i^{no-int}(t)\left(1 - \text{sinc}^2\left[\frac{D(x_i - x_k)}{\lambda L}\right] \text{sinc}^2\left[\frac{D(y_i - y_k)}{\lambda L}\right]\right), \end{aligned} \quad (5.61)$$

where $P_i^{no-int}(t)$ is the power that can be received when user i has no other active user within one beamwidth, and thus no need for interference suppression, with the assumption of the same amount of transmit power used. In particular, when active users are very close, the desired signal power also approaches zero, i.e., as $(x_k, y_k) \to (x_i, y_i)$, $P_i(x_i, y_i, t) \to 0$.

Next, we consider the case of optimum suppression which does not necessarily achieve zero interference, but maximizes the throughput in (5.37). As explained before, non-zero interference $U_i(x_k, y_k, t) \neq 0$ leads to $\gamma_{ik}(t) = 0$ and $\frac{\partial C_k}{\partial V_i^*} \neq 0$. By definition, the optimum suppression scheme outperforms complete cancellation all the time. In some cases, the performance of complete cancellation is very close to that of optimum suppression, and interference is almost completely suppressed though it still has $U_i(x_k, y_k, t) \neq 0$. From (5.54) we have

$$\mu(\xi, \eta, t) V_i(\xi, \eta, t) = \frac{\partial f}{\partial C_i}\frac{\partial C_i}{\partial V_i^*} + \sum_{k \neq i} \frac{\partial f}{\partial C_k}\frac{\partial C_k}{\partial V_i^*}, \quad (5.62)$$

where

$$\frac{\partial C_i}{\partial V_i^*} = \frac{W U_i(x_i, y_i, t)}{W N_0 + \sum_h \alpha_i^2 P_h(x_i, y_i, t)} \cdot \frac{\alpha_i^2}{\lambda L} e^{j\frac{2\pi}{\lambda L}(x_i \xi + y_i \eta)} \quad (5.63)$$

and

$$\frac{\partial C_k}{\partial V_i^*} = -\frac{W U_i(x_k, y_k, t)}{W N_0 + \sum_h \alpha_k^2 P_h(x_k, y_k, t)} \cdot \frac{\alpha_k^2 P_k(x_k, y_k, t)}{W N_0 + \sum_{h \neq k} \alpha_k^2 P_h(x_k, y_k, t)} \cdot \frac{\alpha_k^2}{\lambda L} e^{j\frac{2\pi}{\lambda L}(x_k \xi + y_k \eta)}. \quad (5.64)$$

The term of $\frac{\partial f}{\partial C_k}\frac{\partial C_k}{\partial V_i^*}$ adjusts the phase that has a component of a directional vector (x_k, y_k), and optimally suppresses the interference of the i^{th} signal to the k^{th} user.

In most cases, there is no closed-form solution. Instead, numerical answers can be obtained.

Under severe interference, SDM with interference suppression does not perform well because the desired signal is suppressed too much as well. The exact capacity in this situation, which is called a "Gaussian interference channel [14]," is still an open problem. In a two-user Gaussian interference channel (Fig. 5-3) user 1 and 2 receive signals U_1 and U_2 respectively from transmitted information V_1 and V_2, given as

$$U_1 = \alpha_1(V_1 + cV_2) + Z_1 \tag{5.65}$$

and

$$U_2 = \alpha_2(V_2 + cV_1) + Z_2, \tag{5.66}$$

where α_1 and α_2 represent signal attenuation due to weather conditions ($0 \leq \alpha_1, \alpha_2 \leq 1$), c (> 0) is a scaling factor for symmetric interference, and Z_1 and Z_2 are independent additive white Gaussian noises (AWGN). Our case is when $c < 1$, i.e., the interference is weaker than the desired signal.[4] So far, orthogonal schemes such as time or frequency division multiplexing (TDM or FDM) are known to give the best performance when interference is less than the desired signal power but larger than some amount for $c^* < c < 1$ with some constant c^* [12]. Thus, when the active users are very close-in and suffer severe interference, sequential service in a form of TDM outperforms any interference suppression scheme, by providing orthogonal signals to users. For $0 < c \leq c^*$, SDM with interference suppression is better than sequential service. The threshold c^* is decided by comparing interference suppression and sequential service, which is shown in the following numerical examples.

For a simple example, suppose that M_{act} active users are uniformly located on a line and the nearest neighbors are separated by distance l. The interference between users and thus, their capacities are functions of l. Different l will lead to different antenna patterning. All active users are assumed to have identical static conditions

[4]With strong interference of $c \geq 1$, it has been shown that the receivers can achieve the same capacity region as with no interference of $c = 0$ [27, 50].

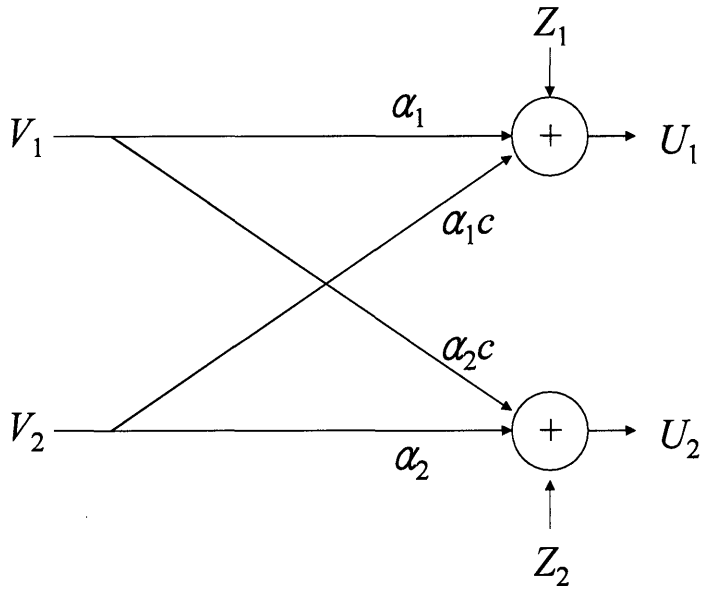

Figure 5-3: A two-user Gaussian interference channel

of average arrival rate A and signal attenuation α^2 over the time interval of interest, which will result in identical antenna gain patterning and power/beam allocation for all the users. We just look at the maximally achievable capacity of one user. Without loss of generality, we solve the antenna gain patterning for user 0 located at $x = 0$ as considering interference $U_0(kl)$ to other $M_{act} - 1$ users located at $x = kl$ for $k = 1, \cdots, M_{act} - 1$ within one side of a mainlobe, i.e., $(M_{act} - 1)l < \frac{\lambda L}{D}$. Here we consider a linear antenna of $|\xi| \leq \frac{D}{2}$, and the result can be easily extended to a planar antenna. With the complete cancellation scheme in (5.57), we have

$$V_0(\xi) = \frac{1 + \sum \gamma_k e^{j\frac{2\pi}{\lambda L} kl\xi}}{\mu'}, \tag{5.67}$$

where $\gamma_k \equiv \frac{\gamma_{0k}}{s_0}$ and $\mu' \equiv \frac{\mu}{s_0}$. The same amount of power is allocated to each of M_{act} active users with $|V_i(\xi)|^2 = \frac{\rho_0}{M_{act}}$ for $i = 0, \cdots, M_{act} - 1$. Then, we have

$$U_0(x) = \frac{D}{\lambda L}\frac{1}{\mu'}\left\{\text{sinc}\left(\frac{xD}{\lambda L}\right) + \sum_{k=1}^{M_{act}-1} \gamma_k \text{sinc}\left[\frac{(x - kl)D}{\lambda L}\right]\right\} \tag{5.68}$$

with $U_0(il) = 0$ for $i = 1, \cdots, M_{act} - 1$. We obtain γ_k by solving a set of linear equations, given as

$$-\sum_{k=1}^{M_{act}-1} \gamma_k \text{sinc}\left[(i-k)\frac{lD}{\lambda L}\right] = \text{sinc}\left[i\frac{lD}{\lambda L}\right] \tag{5.69}$$

for $i = 1, \cdots, M_{act} - 1$. μ' is determined such that $|V_0(\xi)|^2 = \frac{\rho_0}{M_{act}}$, given as

$$\mu'^2 = \frac{M_{act}}{\rho_0}\left\{1 + \sum_k \gamma_k^2 - 2\sum_k \gamma_k \text{sinc}\left[\frac{klD}{\lambda L}\right] + 2\sum_k \sum_{i>k} \gamma_k \gamma_i \text{sinc}\left[(k-i)\frac{lD}{\lambda L}\right]\right\}. \tag{5.70}$$

With the optimum suppression scheme in (5.62), we have

$$V_0(\xi) = \frac{1 + \sum \psi_k e^{j\frac{2\pi}{\lambda L}kl\xi}}{\mu'} \tag{5.71}$$

with

$$\psi_k = -\frac{\alpha^2 U_0(0) \cdot U_0(kl)}{WN_0 + \alpha^2 U_0^2(kl)}. \tag{5.72}$$

and

$$\mu'^2 = \frac{M_{act}}{\rho_0}\left\{1 + \sum_k \psi_k^2 - 2\sum_k \psi_k \text{sinc}\left[\frac{klD}{\lambda L}\right] + 2\sum_k \sum_{i>k} \psi_k \psi_i \text{sinc}\left[(k-i)\frac{lD}{\lambda L}\right]\right\}. \tag{5.73}$$

Note that we have symmetric

$$U_0(0) = U_k(kl) \quad \text{and} \quad U_0(kl) = U_k(0) \tag{5.74}$$

for $k = 1, \cdots, M_{act} - 1$ by assuming identical static conditions for all active users. We obtain ψ_k numerically in terms of $U_0(0)$ and $U_0(kl)$, which are given as

$$U_0(0) = \frac{D}{\lambda L}\frac{1}{\mu'}\left\{1 - \sum_k \psi_k \text{sinc}\left(\frac{klD}{\lambda L}\right)\right\} \tag{5.75}$$

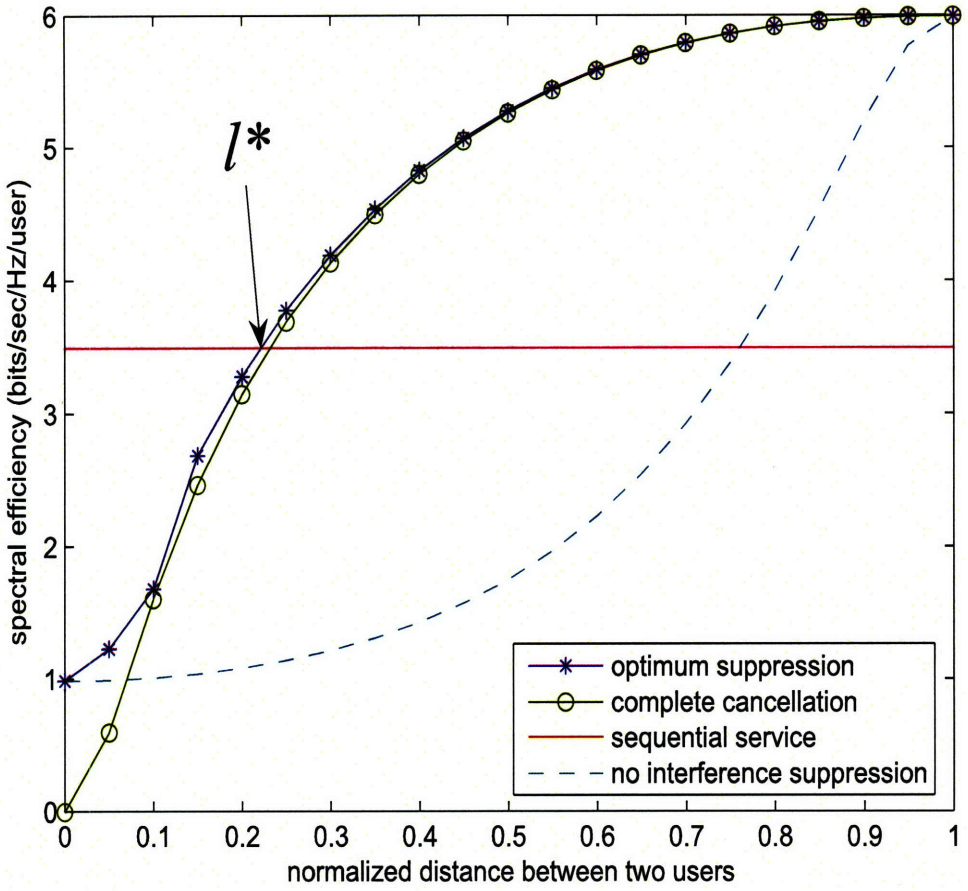

Figure 5-4: Capacity of one user as a function of the distance (normalized by one beamwidth) between two users in high SNR of $\frac{E_b}{N_0} = 10.2$ dB

and

$$U_0(kl) = \frac{D}{\lambda L}\frac{1}{\mu'}\left\{\text{sinc}\left(\frac{klD}{\lambda L}\right) - \sum_m \psi_m \text{sinc}\left[\frac{(m-k)lD}{\lambda L}\right]\right\}. \tag{5.76}$$

As l approaches zero, all of the desired signal and interference approach to $\frac{D}{\lambda L}\sqrt{\frac{\rho_0}{M_{act}}}$.

We compare capacities of two schemes of complete cancellation in (5.57) and optimum suppression in (5.62), by changing the distance between two active users ($M_{act} = 2$). The scheme without interference suppression in (5.47) is also shown as a benchmark with different SNRs (without considering interference) in Fig. 5-4 and 5-5.

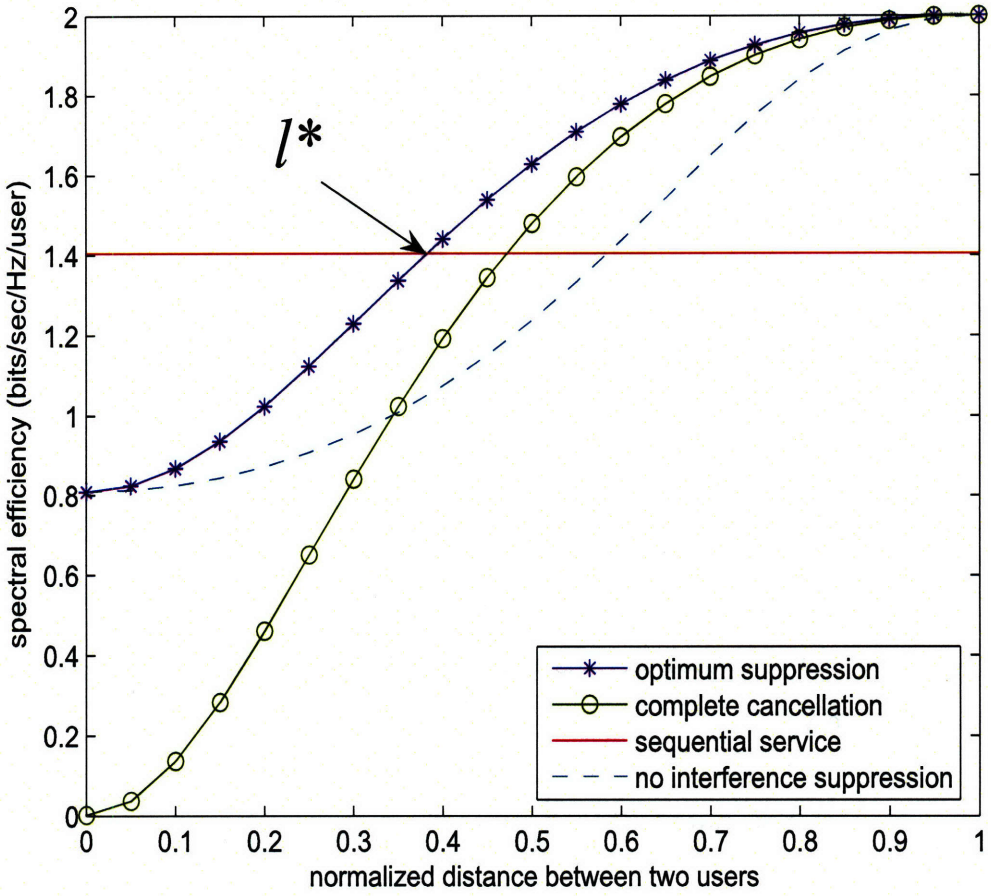

Figure 5-5: Capacity of one user as a function of the distance (normalized by one beamwidth) between two users in low SNR of $\frac{E_b}{N_0} = 1.76$ dB

The scheme with no interference suppression results in interference of

$$U_0(l) = U_1(0) = \frac{D}{\lambda L}\sqrt{\frac{\rho_0}{2}}\text{sinc}\left(\frac{lD}{\lambda L}\right) \tag{5.77}$$

and desired signal power of

$$U_0(0) = U_1(l) = \frac{D}{\lambda L}\sqrt{\frac{\rho_0}{2}}. \tag{5.78}$$

The complete cancellation scheme with two users has

$$\gamma_1 = -\text{sinc}\left(\frac{lD}{\lambda L}\right) \tag{5.79}$$

and

$$\mu'^2 = \frac{2}{\rho_0}\left[1 - \text{sinc}^2\left(\frac{lD}{\lambda L}\right)\right], \tag{5.80}$$

which leads to

$$U_0(0) = \frac{D}{\lambda L}\sqrt{\frac{\rho_0}{2}}\left[1 - \text{sinc}^2\left(\frac{lD}{\lambda L}\right)\right]. \tag{5.81}$$

We note that

$$U_0(0) \to 0 \quad \text{as} \quad l \to 0, \tag{5.82}$$

which is the biggest drawback of the complete cancellation scheme that also suppresses the desired signal power as the active users are very close to each other. The optimum suppression scheme with

$$\psi_1 \equiv \psi \quad \text{and} \quad \mu'^2 = \frac{2}{\rho_0}\left[1 + \psi^2 - 2\psi\text{sinc}\left(\frac{lD}{\lambda L}\right)\right] \tag{5.83}$$

has the desired signal and interference of

$$U_0(0) = \frac{D}{\lambda L}\sqrt{\frac{\rho_0}{2}}\frac{1 - \psi\text{sinc}\left(\frac{lD}{\lambda L}\right)}{\sqrt{1 + \psi^2 - 2\psi\text{sinc}\left(\frac{lD}{\lambda L}\right)}} \tag{5.84}$$

and

$$U_0(l) = U_1(0) = \frac{D}{\lambda L}\sqrt{\frac{\rho_0}{2}}\frac{\text{sinc}\left(\frac{lD}{\lambda L}\right) - \psi}{\sqrt{1 + \psi^2 - 2\psi\text{sinc}\left(\frac{lD}{\lambda L}\right)}}. \tag{5.85}$$

As l approaches zero, both of desired signal and interference approach to $\frac{D}{\lambda L}\sqrt{\frac{\rho_0}{2}}$, which is the same result as in the scheme without interference suppression.

In the high SNR region (Fig. 5-4) of $\frac{E_b}{N_0} = 10.2$ dB, where E_b is the average signal energy per bit and N_0 is the noise power, the gap between optimum suppression and complete cancellation is very small, so that complete cancellation can be a good

approximation to optimum suppression except when the distance l between two users is extremely close, $l < 0.1 \cdot \frac{\lambda L}{D}$. On the other hand, in the low SNR region (Fig. 5-5) of $\frac{E_b}{N_0} = 1.76$ dB, the gap between two schemes is wider for every distance l and the complete cancellation scheme is even worse than the scheme without interference suppression for a wide range of distances, $l < 0.35 \cdot \frac{\lambda L}{D}$. If active users are closer than some threshold of l^* that is decided by comparing interference suppression and sequential service ($l^* = 0.22 \cdot \frac{\lambda L}{D}$ in the high SNR of Fig. 5-4 and $l^* = 0.38 \cdot \frac{\lambda L}{D}$ in the low SNR of Fig. 5-5), signal degradation is too severe due to co-channel interference and it is better to provide sequential service. At $l^* < l < \frac{\lambda L}{D}$, active users share the bandwidth and timeslots, and appropriate optimum antenna gain patterning is deployed with optimally suppressed interference depending on the operating SNR level. When we have to schedule multiple active users within $\frac{\lambda L}{D}$ at the same time, whether we use interference suppression or sequential service, we lose some spectral efficiency, which can be more than 40 % of the no-interference case (of sparsely located users) for the high SNR example of Fig. 5-4.

In practice, current satellite communication systems operate at the spectral efficiency of $1 \sim 2$ bits/sec/Hz, e.g., by the use of binary or quadrature phase shift keying (BPSK or QPSK) modulation. Since the bandwidth is precious in the satellite communications, bandwidth efficient modulation (BEM) such as M-ary quadrature amplitude modulation (QAM) with $M = 16, 64, \cdots$ is considered for a use in advanced satellite systems. Nonlinearity of the satellite channel makes the use of multi-layered envelope modulation scheme difficult. Moreover, the linear increase of spectral efficiency with a high-order modulation scheme requires the near-exponential increase of SNR, according to the Shannon limit, given as

$$\frac{R}{W} \leq \log_2\left(1 + \frac{R}{W} \cdot \frac{E_b}{N_0}\right) \tag{5.86}$$

$$\Longrightarrow \frac{E_b}{N_0} \geq \frac{2^{\frac{R}{W}} - 1}{R/W} \tag{5.87}$$

for reliable transmission with bit rate $R \leq C$. Even with powerful coding schemes such

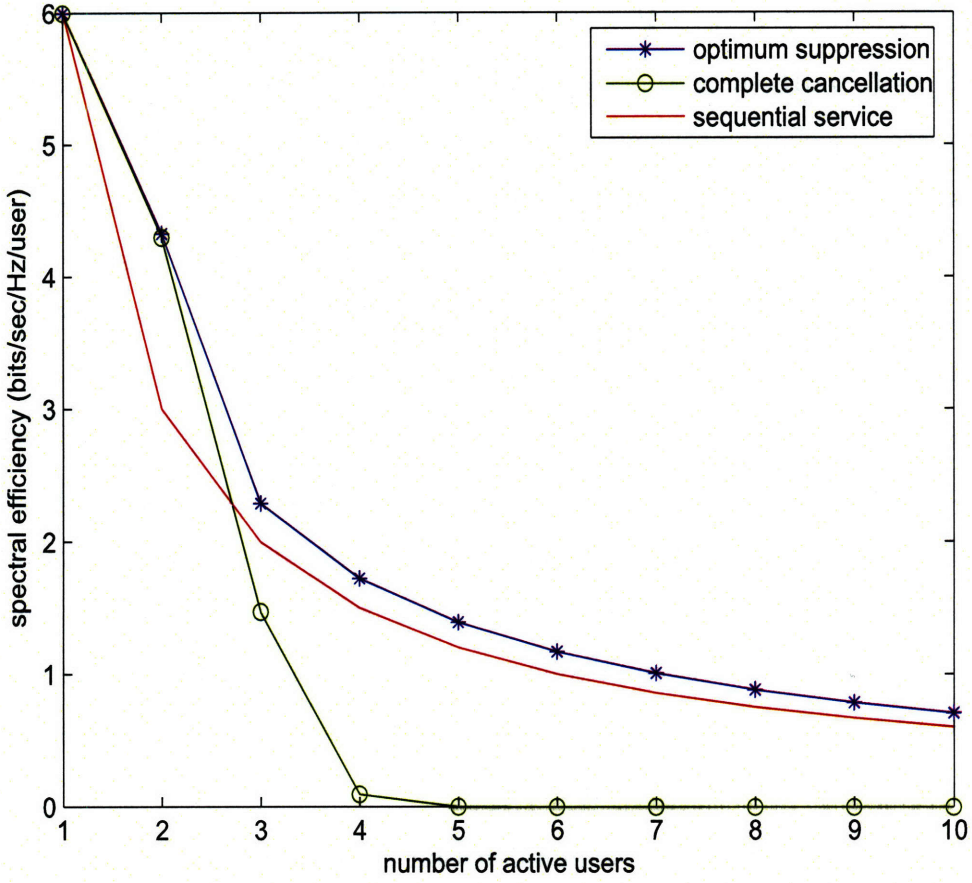

Figure 5-6: Capacity of one user as a function of the number of active users within one beamwidth in high SNR of $\frac{E_b}{N_0} = 10.2$ dB

as Turbo codes and low density parity check (LDPC) codes, huge power consumption makes it almost infeasible to apply a high-order modulation scheme for a TWTA-based satellite system. A future satellite system will be designed to provide better spectral efficiency as high as 6 bits/sec/Hz as in Fig. 5-4 by the use of phased array antenna and SSPAs that have a wide range of linearity and power allocation flexibility.

Fig. 5-6 and 5-7 show the performance comparison of different schemes in terms of the number of uniformly located active users ($1 \leq M_{act} \leq 10$) within $0 < l \leq \frac{\lambda L}{D}$ for high and low SNR levels as in Fig. 5-4 and 5-5. We observe the advantage of multiple signals over a single beam of sequential service for a small number of active

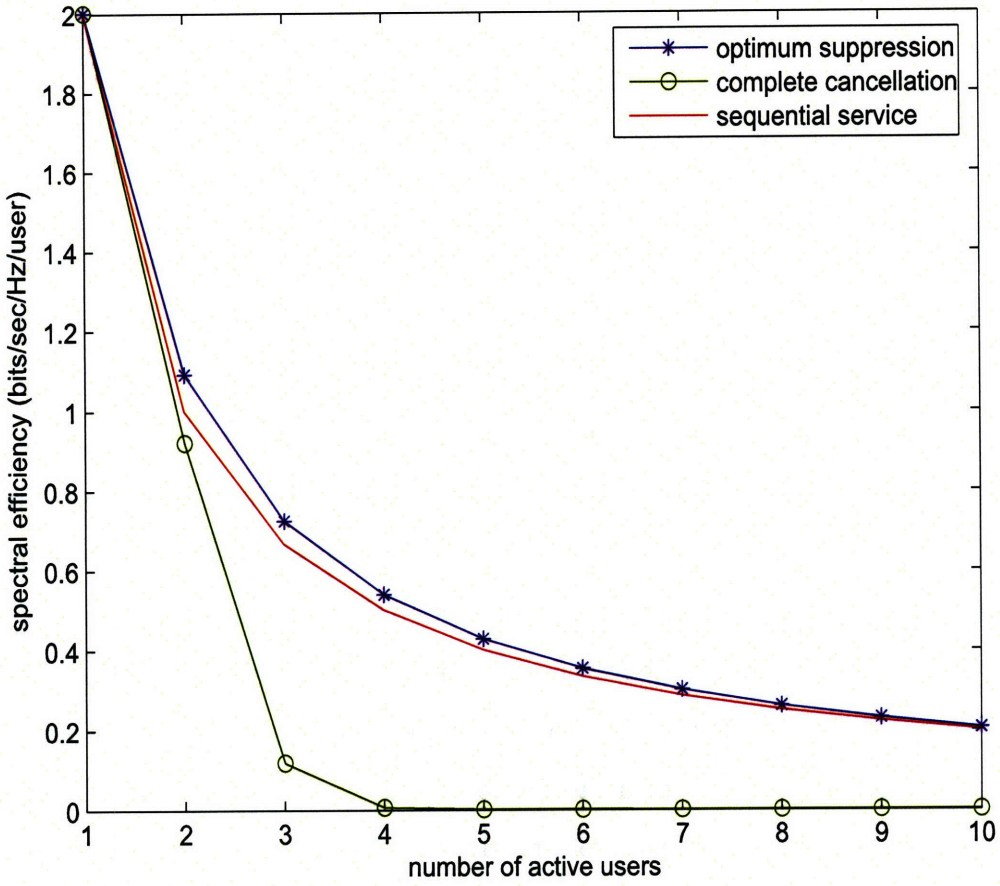

Figure 5-7: Capacity of one user as a function of the number of active users within one beamwidth in low SNR of $\frac{E_b}{N_0} = 1.76$ dB

users $M_{act} = 2$ or 3, in spite of the power loss from interference suppression. The complete cancellation scheme is very vulnerable to multiple active users of $M_{act} \geq 3$. As the number of active users increases, the gap between optimum suppression and sequential service decreases and is negligible. Eventually, sequential service outperforms optimum suppression for a large number of active users (not shown in the figures), which is consistent with what we have seen for the Gaussian interference channel problem and our two-user example in Fig. 5-4 and 5-5: sequential service is better than SDM (with or without interference suppression) under severe interference. Thus, the use of simple sequential service can be recommended when an area is

crowded with many active users, which will be validated in the next section of beam scheduling.

5.3 Beam Scheduling

We now address the problem of beam scheduling with consideration of co-channel interference between close-in active users and the corresponding gain patterning for interference suppression. For the given users that are widely spread over the satellite coverage area, we already showed that it is optimum to pattern the narrowest spot-beams. We will also prove the optimality of scheduling the narrowest spotbeams for the maximum throughput when we can choose active users sparsely located. Fig. 5-4 shows that in the high SNR region, higher than 10 dB for bandwidth efficiency of 6 bits/sec/Hz or beyond that is good for an advanced future communication satellite, complete cancellation of interference is a close approximation to optimum suppression until sequential service outperforms as active users become closer. Even in low SNR, complete cancellation can achieve more than 90 % of optimum suppression and be a reasonable approximation.

We thus model the capacity of the i^{th} user in service, given as

$$C_i = W \log\left(1 + \frac{\alpha_i^2 H_i P_i}{W N_0}\right), \tag{5.88}$$

where H_i (> 0) represents the power loss by deploying complete cancellation and is determined by the distance to other active users. P_i, same as P_i^{no-int} in (5.61), is the received power when user i has no interference suppression and can be calculated from the transmit power by (5.50). If active user i has another active user k at distance $l < \frac{\lambda L}{D}$, the complete cancellation scheme gives

$$H_i = 1 - \mathrm{sinc}^2\left(\frac{l}{\lambda L/D}\right), \tag{5.89}$$

which is derived from (5.61) when $x_i - x_k = l$ and $y_i - y_k = 0$. If $l \geq \frac{\lambda L}{D}$, we ignore

interference and have $H_i = 1$. Note that H_i depends only on the distance to other active user, not on other parameters such as SNR and demand. Here, the value of H_i only considers the nearest active user and it is assumed that the square transmit antenna can be aligned toward the nearest active user to give the best performance. We see that

$$1 - \text{sinc}^2\left(\frac{Dl_x}{\lambda L}\right)\text{sinc}^2\left(\frac{Dl_y}{\lambda L}\right) \leq 1 - \text{sinc}^2\left(\frac{Dl}{\lambda L}\right) \tag{5.90}$$

with $l^2 = l_x^2 + l_y^2$. This approximation and assumption are reasonable because we do not provide antenna gain patterning with interference suppression for more than 3 active users in a crowded area as shown in previous examples. One advantage of using this model is to decouple different signals and their capacities with respect to a set of transmit power $\{P_i\}_{i=1}^M$.

The phased array antenna can cycle very rapidly among the users. Thus, by changing beam allocation variable $z_i(t)$ much faster than average delay deadlines Δ_i, we can serve the back-logged and newly arrived packets of all M users and can maximize $\theta(t)$, and thus the throughput. The question is how the active users should be chosen and clustered each time to maximize the throughput. We now solve the optimum scheduling problem with the time-varying congestion control variable $\theta(t)$. The problem is restated as

$$\text{maximize} \quad \theta(t) \tag{5.91}$$

$$\text{subject to} \quad \bar{d}_i(t) \leq \Delta_i \tag{5.92}$$

$$\sum_{i=1}^{M} z_i(t) \leq K \quad \text{with } z_i(t) = 0 \text{ or } 1 \tag{5.93}$$

$$\text{and} \quad \sum_{i=1}^{M} P_i(t) \leq P_{total}. \tag{5.94}$$

We have average delay and maximum K signal constraints every time t in (5.92) and (5.93). A simple form of the power constraint in (5.94) is derived by applying the average power density constraint of (5.35) to the entire aperture as in (5.50). The power loss by antenna patterning other than narrowest spotbeams is considered in

$H_i < 1$. Since it is hard to solve the binary problem, we relax the binary constraint of $z_i(t)$ and instead apply the Kuhn-Tucker condition, to see which users should be served, by observing $P_i(t)$. The corresponding Lagrangian function (with time index omitted for simplicity) is

$$J(P_i) = \theta - \sum \kappa_i \cdot (\bar{d}_i - \Delta_i) - \Lambda \cdot \left(\sum P_i - P_{total}\right) - \sum \nu_i \cdot (-P_i), \tag{5.95}$$

where we have Lagrangian variables $\kappa_i \geq 0$ for the average delay constraints, Λ (≥ 0) for the total power constraint and ν_i for $-P_i \leq 0$. Differentiating with respect to P_i gives

$$\frac{\partial J}{\partial P_i} = \frac{\partial \theta}{\partial P_i} + \kappa_i \left|\frac{\partial \bar{d}_i}{\partial P_i}\right| - \Lambda + \nu_i = 0, \tag{5.96}$$

where we use

$$\frac{\partial \bar{d}_i}{\partial P_i} = -\left|\frac{\partial \bar{d}_i}{\partial P_i}\right|. \tag{5.97}$$

When we have the optimum $P_i^* > 0$, we have $\nu_i = 0$ and

$$\left.\frac{\partial \theta}{\partial P_i}\right|_{P_i=P_i^*} + \kappa_i \left|\frac{\partial \bar{d}_i}{\partial P_i}\right|_{P_i=P_i^*} = \Lambda < \left.\frac{\partial \theta}{\partial P_i}\right|_{P_i=0} + \kappa_i \left|\frac{\partial \bar{d}_i}{\partial P_i}\right|_{P_i=0}, \tag{5.98}$$

where the concavity of throughput θ and average delay $\bar{d}_i$ in terms of P_i is used as general utility functions are concave.

When $P_j = 0$, we have $\nu_j \geq 0$ and

$$\left.\frac{\partial \theta}{\partial P_j}\right|_{P_j=0} + \kappa_j \left|\frac{\partial \bar{d}_j}{\partial P_j}\right|_{P_j=0} + \nu_j = \Lambda \geq \left.\frac{\partial \theta}{\partial P_j}\right|_{P_j=0} + \kappa_j \left|\frac{\partial \bar{d}_j}{\partial P_j}\right|_{P_j=0}. \tag{5.99}$$

From (5.98) and (5.99), the optimum policy serves K users with the highest value of

$$\left.\frac{\partial \theta}{\partial P_i}\right|_{P_i=0} + \kappa_i \left|\frac{\partial \bar{d}_i}{\partial P_i}\right|_{P_i=0} = \alpha_i^2 H_i \left[\left.\frac{\partial \theta}{\partial C_i}\right|_{C_i=0} + \kappa_i \left|\frac{\partial \bar{d}_i}{\partial C_i}\right|_{C_i=0}\right]. \tag{5.100}$$

The result implies that we have to select better channel conditions with higher α_i^2, less interference with higher H_i and higher marginal returns of the composite cost

function (of the throughput with delay penalties), $f = \theta - \sum \kappa_i \cdot (\bar{d}_i - \Delta_i)$, in terms of allocated capacity with higher $\frac{\partial f}{\partial C_i}\Big|_{C_i=0}$.

To see that the performance of beam scheduling depends on user locations, we consider two extreme cases: (i) for widely spread users, we provide K orthogonal spotbeams, but (ii) for very close-in users, the sequential service by a single beam is the optimum. If the assumption of widely spread user distribution holds, we select K users with highest $\alpha_i^2 \frac{\partial f}{\partial C_i}\Big|_{C_i=0}$ and have $H_i = 1$ for them by serving only one user within one beamwidth with the narrowest spotbeam. The K orthogonal beams provide a form of space division multiplexing (SDM). At next time slot, another set of K users with highest $\alpha_i^2 \frac{\partial f}{\partial C_i}\Big|_{C_i=0}$ and $H_i = 1$ are selected. The cost function f has a penalty for violating average delay constraints, so that it is time-varying according to power/signal allocation, back-logged data and corresponding average delays. The selection process is iterated with different K users and the beams cycle rapidly through the cluster of users. In case of very close-in users, H_i becomes very small when more than one users are served at the same time, and it is better to serve sequentially by using a single beam with all the power. Between these extreme cases, we cannot pick K non-interfering users and have to compare interference suppression and sequential service (with a small number of transmitters on) as a singular case. Close-in users in a good channel condition may receive signals simultaneously if their $\alpha_i^2 H_i$'s are higher than α_j^2 in a worse channel condition. We note that by moving up to the higher frequency band and/or increasing the transmit antenna size, we can make a beamwidth narrower and approach the optimum throughput. Though we solved two separate subproblems of antenna gain patterning and scheduling, the optimum solution suggests that the two designs should be made jointly since the selection of K users and the power loss H_i from gain patterning depend on each other. This is in fact the hardest part to solve in practice and a simple but near-optimum algorithm will be proposed in Section 5.5.

We consider a simple example, where 100 ($= M$) users are uniformly located on a planar grid with distance l between adjacent users. We assume identical traffic loads

and channel conditions for users. The total area occupied by M users is approximately Ml^2 and the total coverage area of K orthogonal beams is $K(\frac{\lambda L}{D})^2$ because adjacent beams need to be at least $\frac{\lambda L}{D}$ apart for negligible interference. Thus, if $Ml^2 \geq K(\frac{\lambda L}{D})^2$, each user has average beam allocation of

$$\bar{z_i} = \frac{K}{M}, \tag{5.101}$$

which is equal to 0.2 in this example with $K = 20$, and we can schedule K orthogonal beams all the time with

$$P_i = \frac{P_{total}}{K} \quad \text{and} \quad H_i = 1 \tag{5.102}$$

for every i in service. In fact, if $l < \frac{\lambda L}{D}$, one may not be able to schedule perfect K orthogonal beams all the time, and will lose some capacity due to interference itself or interference suppression. We will suppress this exact analysis until the simulation result part, which is given in Section 5.6, and assume perfect scheduling for $Ml^2 > K(\frac{\lambda L}{D})^2$.

We now compare interference suppression (with K transmitters on) and sequential service (with less than K transmitters on) when $Ml^2 < K(\frac{\lambda L}{D})^2$. If K signals are transmitted when $Ml^2 = K(\frac{\lambda L/D}{M_{act}})^2$ with $M_{act} = 2, 3, \cdots$, the distance between adjacent beams is $\frac{\lambda L/D}{M_{act}}$ and there are M_{act} active users within one beamwidth in each direction of x and y in the planar grid. In the same way as in Fig. 5-4 and 5-5, we can calculate capacity per user with optimum suppression or complete cancellation of interference, which is plotted in Fig 5-8. When using sequential service for close-in users, we need K' $(< K)$ orthogonal beams such that $K'(\frac{\lambda L}{D})^2 = Ml^2$. We set an integer K', given as

$$K' = \max\left\{1, \left\lfloor \frac{Ml^2}{(\lambda L/D)^2} \right\rfloor \right\}, \tag{5.103}$$

where $\lfloor x \rfloor$ is a floor function. With K' $(< K)$ beams, average beam allocation $\bar{z_i}$ decreases from $\frac{K}{M}$ to $\frac{K'}{M}$ but allocated power increases from P_{total}/K to P_{total}/K'. Both interference suppression schemes outperform sequential service over wide range of distance except for very close-in users. The beam scheduling choice of interference

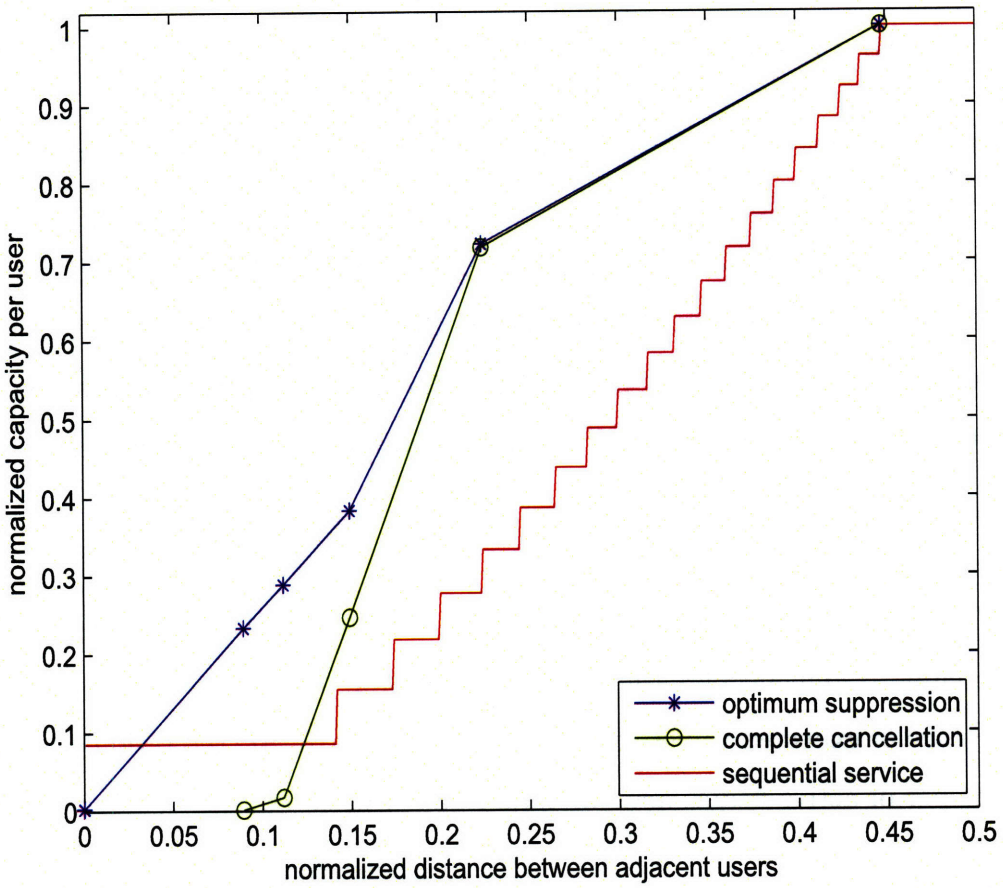

Figure 5-8: Normalized capacity per user as a function of the distance (normalized by one beamwidth) in the example of uniformly located users with an identical amount of traffic for each

suppression or sequential service depends on user distribution and distance between users. The plot of sequential service resembles a "stair" as the number of orthogonal beams used decreases one by one from $K = 20$ to $K' = 1$.

5.4 Comparison of Phased Array Antenna and Multiple Beam Antenna

Here we compare the average steady-state performance of the phased array antenna with that of the multiple beam antenna, which was developed in Chapter 4, suppressing the issue of interference (with $H_i = 1$ for every i). While the multiple beam antenna has a power constraint for each beam, the phased array antenna can provide any power level (up to the total power) for a signal. In Chapter 4, it is shown that the multiple beam antenna provides different performance levels in the following two cases:

1. When there is no dominant user (with $\bar{z}_i < 1$ for every i), the multiple beam antenna system decides the congestion control parameter θ in terms of the average of expected values of user parameters and K/M, the ratio between the number of active beams and the total number of users.

2. When there is one dominant user, the multiple beam antenna provides one whole active beam $\bar{z}_i = 1$ for the dominant user, but under-utilizes resources for other users.

In the following, we show that due to its flexibility for power allocation the phased array antenna can give higher throughput (i.e., θ) in the second case, by increasing the average power for the dominant user.

Let Q_i denote the average power allocated for the i^{th} user when the signal is on, given as,

$$Q_i = E[P_i | z_i = 1]. \tag{5.104}$$

With $\bar{z}_i = \Pr[z_i = 1]$, we have an average power constraint of

$$\sum_i \bar{P}_i = \sum_i \bar{z}_i Q_i \leq P_{total}. \tag{5.105}$$

Furthermore, we assume that the average capacity C_i^{avg} is a simple function of Q_i and $\bar{z}_i$, given as

$$C_i^{avg} = \bar{z}_i \cdot W \log\left(1 + \frac{\alpha_i^2 Q_i}{W N_0}\right) \tag{5.106}$$

as if the signal for the i^{th} user has either $P_i = Q_i$ with the signal on or $P_i = 0$ with the signal off. We note that this approximation is reasonable when the bandwidth W is very broad or $P_i(t)$ does not change much around Q_i when $z_i(t) = 1$. With the multiple beam antenna, power operation is fixed at $P_i(t) = P_0$ when $z_i(t) = 1$, so that $Q_i = P_0$ for every i.

The optimization problem with the modified constraints in terms of $\bar{z}_i$ and Q_i is given as

$$\text{maximize} \quad \theta \tag{5.107}$$

$$\text{subject to} \quad \bar{d}_i \leq \Delta_i \tag{5.108}$$

$$\sum_i \bar{z}_i Q_i \leq P_{total} \tag{5.109}$$

$$\sum_{i=1}^{M} \bar{z}_i \leq K \tag{5.110}$$

$$\text{and} \quad 0 \leq \bar{z}_i \leq 1 \quad \text{for every } i. \tag{5.111}$$

The Lagrangian function is given as

$$J(\bar{z}_i, Q_i) = f - \mu\left(\sum \bar{z}_i Q_i - P_{total}\right) - \Lambda\left(\sum \bar{z}_i - K\right) - \sum \Lambda_i\left(\bar{z}_i - 1\right), \tag{5.112}$$

$$\text{with} \quad f = \theta - \sum \kappa_i \cdot (\bar{d}_i - \Delta_i) \tag{5.113}$$

and differentiated with respect to $\bar{z}_i$ and Q_i, given as

$$\frac{\partial J}{\partial \bar{z}_i} = \frac{\partial f}{\partial C_i^{avg}} \cdot \frac{\partial C_i^{avg}}{\partial \bar{z}_i} - \mu Q_i - \Lambda - \Lambda_i = 0 \tag{5.114}$$

and

$$\frac{\partial J}{\partial Q_i} = \frac{\partial f}{\partial C_i^{avg}} \cdot \frac{\partial C_i^{avg}}{\partial Q_i} - \mu = 0. \tag{5.115}$$

Combining (5.114) and (5.115) and replacing C_i^{avg} with (5.106), we obtain

$$\frac{\mu W \log\left(1+\frac{\alpha_i^2 Q_i}{WN_0}\right)}{\frac{\alpha_i^2/N_0}{1+\frac{\alpha_i^2 Q_i}{WN_0}}} - \mu Q_i = \Lambda + \Lambda_i \tag{5.116}$$

Assuming that full power is used with $\mu \neq 0$, we have

$$\left(1+\frac{\alpha_i^2 Q_i}{WN_0}\right)\cdot \log\left(1+\frac{\alpha_i^2 Q_i}{WN_0}\right) - \frac{\alpha_i^2 Q_i}{WN_0} = \alpha_i^2 \cdot (\Lambda' + \Lambda_i'), \tag{5.117}$$

where $\Lambda' = \frac{\Lambda}{\mu W N_0}$ (≥ 0) is decided by the total power and maximum K independent signal constraints and we have $\Lambda_i' = \frac{\Lambda_i}{\mu W N_0} > 0$ only if $\bar{z}_i = 1$ and $\Lambda_i' = 0$ otherwise.

To compare performances of the phased array antenna and the multiple beam antenna, we provide a simple example, where we have

$$A_1 = A_2 = \cdots = A_{M-1} \leq A_M, \tag{5.118}$$

so that $\bar{z}_i < 1$ for $i = 1, \cdots, M-1$ and $\bar{z}_M = 1$. This example models the real applications of military and commercial systems, where in general there are a small number of very demanding users and a large number of users with small demand. From Eq. (5.117) and

$$\sum_{i=1}^{M-1} \bar{z}_i = K - 1 \tag{5.119}$$

with every $\alpha_i^2 = 1$ in the example, $M-1$ users except the M^{th} have the same average power and beam allocation, given as

$$Q_i = Q_0 \quad \text{and} \quad \bar{z}_i = \frac{K-1}{M-1} \tag{5.120}$$

with the phased array antenna. Q_0 and Q_M satisfies

$$\sum_{i=1}^{M} \bar{z}_i Q_i = \sum_{i=1}^{M-1} \bar{z}_i Q_0 + Q_M = (K-1)Q_0 + Q_M = P_{total}. \tag{5.121}$$

From the average delay constraints with $M/M/1$ queue approximation of

$$\bar{d}_i = \frac{1}{C_i^{avg} - \theta A_i} \tag{5.122}$$

and the same average delay deadline Δ for every i, the optimum θ_ϕ of the phased array antenna is given as

$$\begin{aligned} \theta_\phi &= \frac{1}{A_i}\left(C_i^{avg} - \frac{1}{\Delta}\right) = \frac{1}{A_i}\left[\frac{K-1}{M-1} \cdot W \log\left(1 + \frac{Q_0}{WN_0}\right) - \frac{1}{\Delta}\right] \\ &= \frac{1}{A_M}\left(C_M^{avg} - \frac{1}{\Delta}\right) = \frac{1}{A_M}\left[W \log\left(1 + \frac{Q_M}{WN_0}\right) - \frac{1}{\Delta}\right] \end{aligned} \tag{5.123}$$

for $i \neq M$ in the first equality. If we assume $W \to \infty$, we have

$$C_i^{avg} = \frac{K-1}{M-1} \cdot \frac{Q_0}{N_0} \quad \text{for } i \neq M \tag{5.124}$$

and

$$C_M^{avg} = \frac{Q_M}{N_0}, \tag{5.125}$$

which gives

$$\theta_\phi = \frac{1}{\sum^M A_i}\left[(K-1)\frac{Q_0}{N_0} + \frac{Q_M}{N_0} - \frac{M}{\Delta}\right] \tag{5.126}$$

$$= \frac{1}{\frac{1}{M}\sum^M A_i}\left(\frac{1}{M}\frac{P_{total}}{N_0} - \frac{1}{\Delta}\right), \tag{5.127}$$

where we use the following: if

$$r = \frac{X_i}{Y_i} \quad \text{for } i = 1, \cdots, M \tag{5.128}$$

then,

$$r = \frac{\sum_{i=1}^M X_i}{\sum_{i=1}^M Y_i}. \tag{5.129}$$

The performance of the multiple beam antenna is decided by the most demanding user M with

$$\bar{z}_M = 1 \quad \text{and} \quad P_M = P_0 = \frac{P_{total}}{K}, \tag{5.130}$$

which gives

$$\theta_{MBA} = \frac{1}{A_M}\left(\frac{P_{total}}{KN_0} - \frac{1}{\Delta}\right). \tag{5.131}$$

If A_M is not large enough, θ_{MBA} is the same as θ_ϕ because the phased array antenna uses the same Q_i for every i from Eq. (5.117) as the multiple beam antenna does. Otherwise, the asymptotic gain of θ_ϕ over θ_{MBA} with $W \to \infty$ is given as

$$\frac{\theta_\phi}{\theta_{MBA}} = \frac{A_M}{\frac{1}{M}\sum^M A_i} \cdot \frac{\frac{1}{M}\frac{P_{total}}{N_0} - \frac{1}{\Delta}}{\frac{1}{K}\frac{P_{total}}{N_0} - \frac{1}{\Delta}} \tag{5.132}$$

$$\to \frac{A_M}{\frac{1}{M}\sum^M A_i} \cdot \frac{K}{M} \quad \text{if } \Delta \to \infty. \tag{5.133}$$

Thus, the asymptotic gain of the phased array antenna over the multiple beam antenna with very broad bandwidth, large delay and no interference is given as

$$\frac{\theta_\phi}{\theta_{MBA}} = \max\left\{\frac{A_{max}}{\bar{A}} \cdot \frac{K}{M}, 1\right\}, \tag{5.134}$$

where A_{max} is the maximum arrival rate (equal to A_M in this case) and $\bar{A} = \frac{1}{M}\sum_{i=1}^M A_i$ is the average of all the arrival rates A_i, $i = 1, \cdots, M$.

Fig. 5-9 plots θ_ϕ and θ_{MBA} for the above example as A_M/A_0 increases. θ_ϕ is plotted for different $\frac{P_{total}}{KWN_0}$, which shows that the concavity of a capacity function in terms of power reduces θ_ϕ for larger $\frac{P_{total}}{KWN_0}$. With $K = 20$ and $M = 100$, we have $\theta_\phi > \theta_{MBA}$ for $A_M/A_0 > 5$ and the gain of θ_ϕ over θ_{MBA} is near $\frac{A_M}{\bar{A}} \cdot \frac{K}{M}$ with small $\frac{P_{total}}{KWN_0}$. Here we assume one user per beam and and the user is located at the center of the beam for the multiple beam antenna. In practice, the deployment of fixed beams with many users inside makes it hard to focus the center of beam exactly on the user, which is easier to achieve with the phased array antenna.

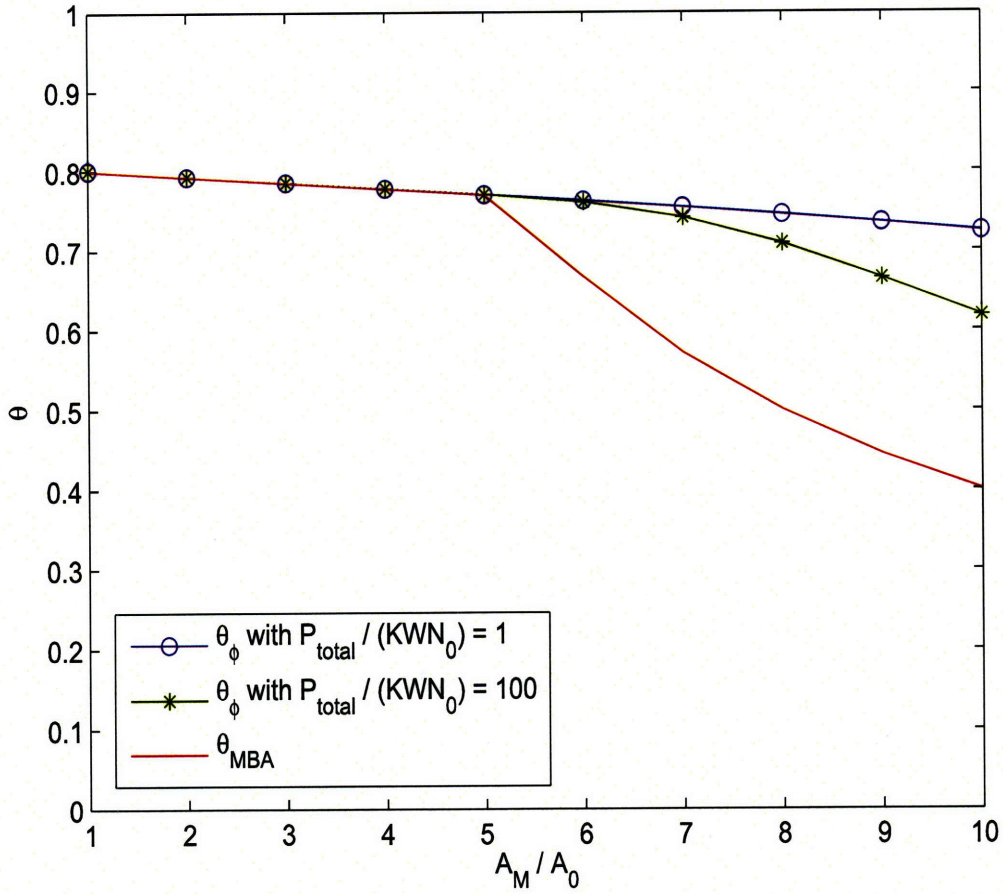

Figure 5-9: Comparison of θ of the phased array antenna and the multiple beam antenna as the demand of one dominant cell increases while the demands of $M-1$ other cells are fixed and uniform

The next example considers a more complicated scenario, where 100 users ($M = 100$) can receive up to 20 signals ($K = 20$) at each timeslot. We assume exponentially distributed incoming traffic of

$$A_i = A_0 \exp(i \cdot \beta), \tag{5.135}$$

where A_0 and β are the parameters that control the shape of traffic distribution. Fig. 5-10 compares θ_ϕ and θ_{MBA} as the traffic distribution (in terms of $A_{max}/\bar{A}$) and the distance between users change. To mitigate interference as the distance decreases,

the multiple beam antenna reduces the number of active beams and the phased array antenna selects the better of interference suppression and sequential service. The result shows that the phased array antenna always performs better than the multiple beam antenna, except when there is no interference and traffic distribution is not extremely unbalanced (which is not shown in the plot). In this case, two antennas give the same performance. The advantage of the phased array antenna over the multiple beam antenna can be shown well especially when traffic distribution is very unbalanced ($A_{max}/\bar{A} > 8$) and/or there is a moderate level of interference between active users, so that interference suppression can be used ($0.2\frac{\lambda L}{D} < l < 0.4\frac{\lambda L}{D}$).

5.5 Near-Optimum Algorithm

We derived an optimum scheduling policy by solving the throughput maximization problem in Section 5.3. Here, from the optimum scheduling result, we develop a near-optimum, low-complexity and real-time algorithm of performing active user selection, antenna gain patterning, power allocation, and admission control.

For the cost function of $f = \theta - \sum \kappa_i(\bar{d}_i - \Delta_i)$, we have

$$\begin{aligned} \left.\frac{\partial f}{\partial C_i}\right|_{C_i=0} &= \left.\frac{\partial \theta}{\partial C_i}\right|_{C_i=0} + \kappa_i \left|\frac{\partial \bar{d}_i}{\partial C_i}\right|_{C_i=0} \\ &= \left.\frac{\partial \theta}{\partial C_i}\right|_{C_i=0} \qquad \text{if } \bar{d}_i < \Delta_i \end{aligned} \tag{5.136}$$

because $\kappa_i > 0$ only if $\bar{d}_i \geq \Delta_i$ by the Kuhn-Tucker condition. Thus, the selection algorithm depends on the admission control policy of the system because $\left.\frac{\partial f}{\partial C_i}\right|_{C_i=0}$ reduces to $\left.\frac{\partial \theta}{\partial C_i}\right|_{C_i=0}$ if the average delay constraint of the i^{th} user is met. We observe that the average delay constraints $\bar{d}_i \leq \Delta_i$ alone cannot guarantee the stability of the system. In real-time applications the measured average queueing delay at the present decreases as more packets are admitted at the cost of packet delay increase in the future and ultimately system instability may occur. Thus, we need another constraint for system stability during the transient state.

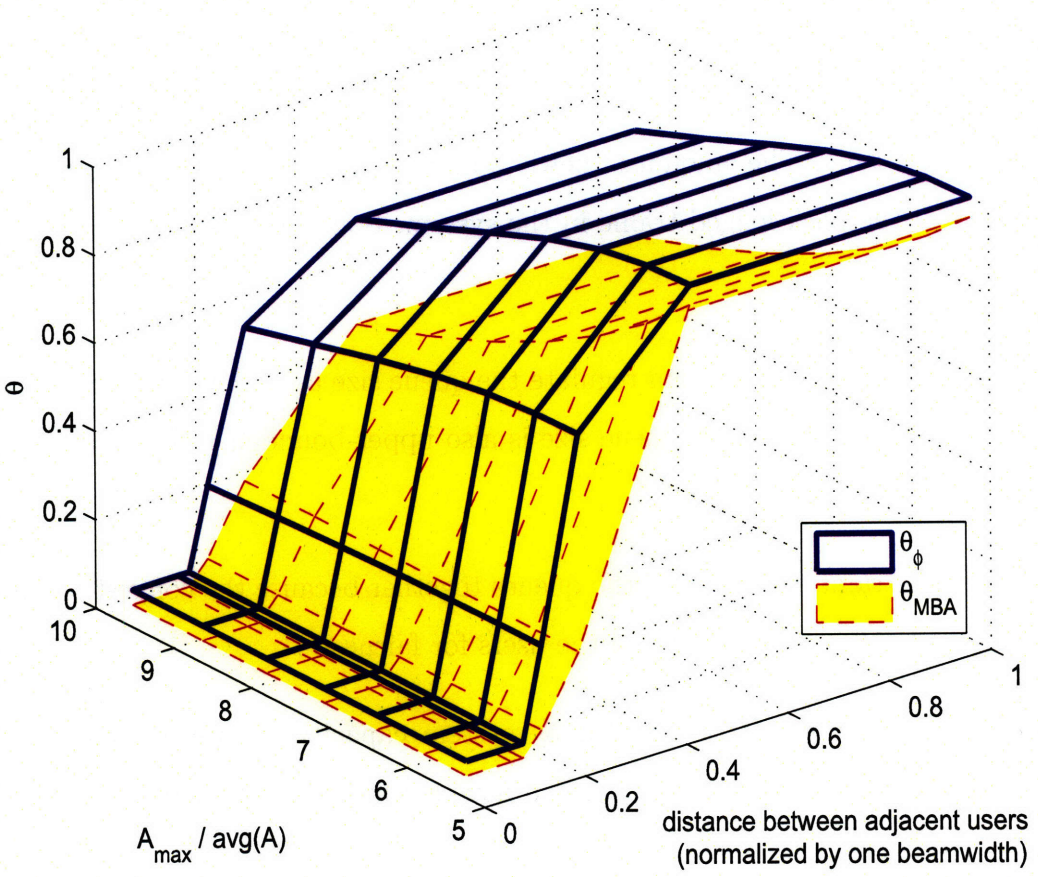

Figure 5-10: Comparison of θ of the phased array antenna and the multiple beam antenna as a function of traffic distribution in terms of $A_{max}/\bar{A}$ and the distance between users that is normalized by one beamwidth

There can be some ways for imposing system stability. Here we propose a constraint of the maximum total of accumulated delays for all the users. Let F_i denote the amount of the accumulated traffic in the i^{th} queue, $B_{i,\tau}$ the amount of traffic having waited for τ timeslots in the i^{th} queue (i.e., the packets were admitted τ timeslots ago, thus the delay is τ and increases every timeslot), and Ψ some constant to upper-bound the total of accumulated delays. The constraint is given as

$$\sum_{i=1}^{M} \bar{d}_i \cdot F_i = \sum_{i=1}^{M} \sum_{\tau=2}^{t} \tau \cdot B_{i,\tau} + 1 \cdot \theta \sum_{i=1}^{M} A_i \leq \Psi, \tag{5.137}$$

where

$$F_i = \sum_{\tau=1}^{t} B_{i,\tau} = \sum_{\tau=2}^{t} B_{i,\tau} + \theta A_i. \tag{5.138}$$

We assume that the new packets just admitted are back-logged immediately with delay 1 ($B_{i,1} = \theta A_i$). The amount of newly admitted traffic is decided by the amount and delays of the back-logged traffic in the system. Some remarks on the constraint are following.

- With the constraint we can regulate the queue size as well as the average delay. Since $1 \leq \bar{d}_i \leq \Delta_i$, each queue size is also upper-bounded.[5] (If $\bar{d}_i = 0$, it is an empty queue.)

- The constraint considers all the queues together because the congestion control parameter θ is universal to all the users for fairness.

- The value of Ψ can be decided by system capacity. If we assume that the maximum queue size is q_{max}, we can have a reasonable range of

$$\Psi < q_{max} \cdot \sum_{i=1}^{M} \Delta_i. \tag{5.139}$$

We calculate $\frac{\partial f}{\partial C_i}\Big|_{C_i=0}$ by the use of constraint (5.137), which is equivalent to

$$\theta \leq \frac{1}{\sum_i A_i} \left[\Psi - \sum_{i=1}^{M} \sum_{\tau=2}^{t} \tau \cdot B_{i,\tau} \right]. \tag{5.140}$$

To maximize θ we have to minimize $\sum_{i=1}^{M} \sum_{\tau=2}^{t} \tau \cdot B_{i,\tau}$, which is to serve the packets with the highest delay as long as the average delay constraint $\bar{d}_i \leq \Delta_i$ is satisfied. When traffic demand is beyond system capacity ($\sum A_i > \Psi$), the equality holds in (5.140) and we have

$$\frac{\partial f}{\partial C_i}\Big|_{C_i=0} = \frac{\partial \theta}{\partial C_i}\Big|_{C_i=0}$$

[5]Every admitted packet has at least a delay of one timeslot in our assumption.

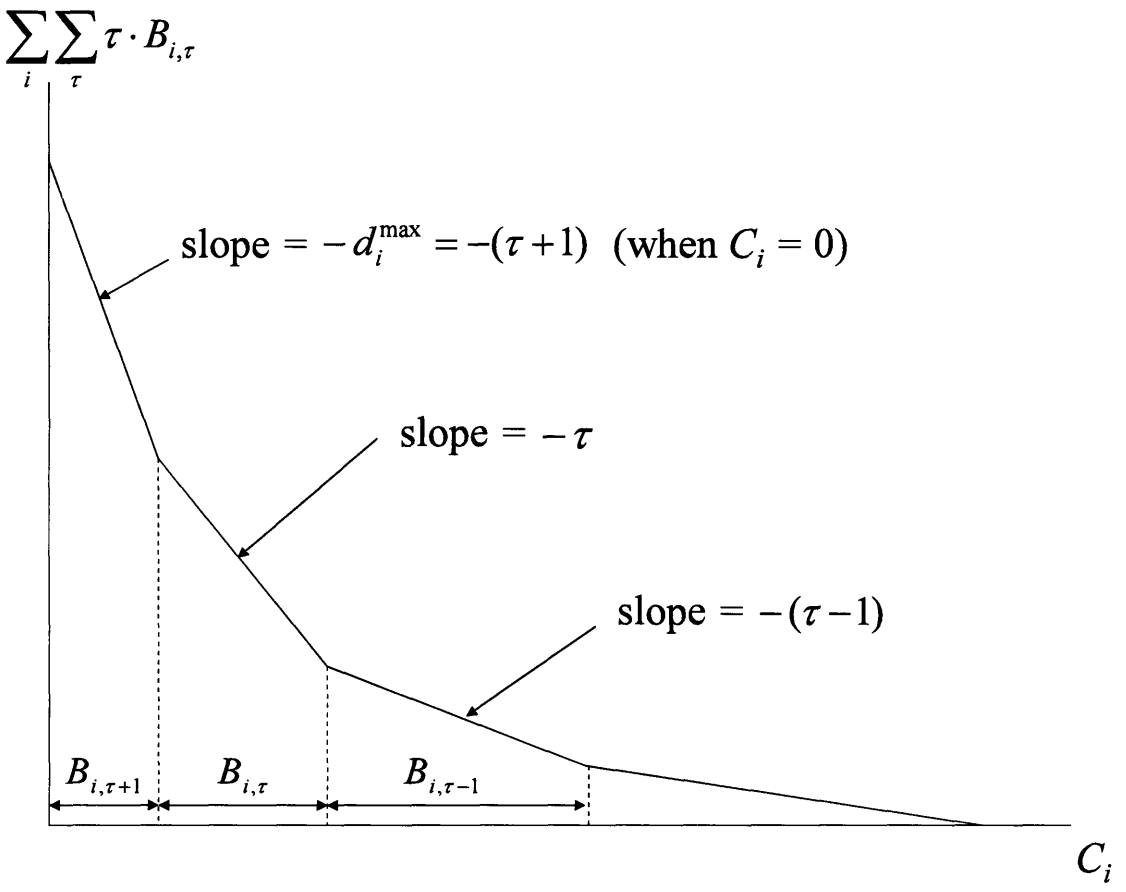

Figure 5-11: A plot of the accumulated delay with respect to the allocated capacity

$$= \frac{-1}{\sum A_i} \frac{\partial}{\partial C_i} \left[\sum_{i=1}^{M} \sum_{\tau=2}^{t} \tau \cdot B_{i,\tau} \right] \Bigg|_{C_i=0}$$
$$= \frac{1}{\sum A_i} d_i^{max}, \tag{5.141}$$

where $d_i^{max} = \max_{\{B_{i,\tau}>0\}} \tau$ is the largest packet delay in the i^{th} queue. By allocating one unit of C_i, we can clear one unit of $B_{i,\tau}$ and it is optimum to serve the packet with the highest delay. Fig. 5-11 plots the change of the accumulated delay, which is piecewise linear and convex, with respect to the allocated capacity. If demand is within system capacity, we can still claim that serving the traffic with the highest delay is a sensible solution, considering the future throughput and delay.

In summary, the selection algorithm, which is the optimum to the maximization

problem (5.91) with an additional constraint (5.137), is to have K active users with (i) $\bar{d}_i = \Delta_i$ and then (ii) largest $\alpha_i^2 H_i d_i^{max}$ (ignoring the universal constant $\frac{1}{\sum A_i}$ to all the users). However, the selection of active users affects the value of H_i and vice versa (joint design of antenna gain patterning and scheduling is reminded), which makes the selection process very complicated and time-consuming. Thus, we suggest a sub-optimum selection algorithm that chooses the most demanding user based on average delay constraints and $\alpha_i^2 d_i^{max}$, and then consider the interference level decided by the users already selected for selecting next active users. The detail of how to select active users is as follows.

1. Select users with $\bar{d}_i = \Delta_i$, and update H_i to remaining users resulted from the selected users.

2. For remaining users (or for all M users if no one is selected in Step 1), look at $\alpha_i^2 H_i d_i^{max}$ and add the user of the biggest value to the list if the user satisfies $l > l^*$, where l is the minimum distance between the user of the biggest value of $\alpha_i^2 H_i d_i^{max}$ and other active users already selected, and l^* is the distance threshold whether sequential service or interference suppression is implemented (in general $0.1\frac{\lambda L}{D} \leq l^* \leq 0.5\frac{\lambda L}{D}$). If $l \leq l^*$, reject the user, proceed to the next user and repeat Step 2.

3. After adding a user, update interference level H_i to remaining users, resulted from the user selected in Step 2.

4. Repeat Step 2 and 3 until either

   i) selecting K active users or

   ii) scanning all M users.

   In case of ii), serve only less than K users.

We allocate the optimum power $P_i^\star$ and the corresponding capacity $C_i^\star$ given as

$$C_i^\star = W \log\left(1 + \frac{\alpha_i^2 H_i P_i^\star}{W N_0}\right), \tag{5.142}$$

to satisfy the followings.

- According to the equation in (5.98), every selected user has an identical marginal return, given as

$$\frac{\alpha_i^2 H_i}{1+\frac{\alpha_i^2 H_i P_i^\star}{WN_0}} \cdot d_i^\star = \Lambda', \tag{5.143}$$

  where Λ' is a Lagrangian multiplier and decided by the total power. $d_i^\star$ is the highest delay after the optimum power allocation, so that we have $d_i^{max}(t+1) = d_i^\star(t)$ in the next timeslot.

- Allocated capacity is equal to the sum of served traffic, given as

$$C_i^\star = \sum_{\tau=d_i^\star+1}^{d_i^{max}} B_{i,\tau} + \tilde{B}_{i,d_i^\star}, \tag{5.144}$$

  where $\tilde{B}_{i,d_i^\star}$ represents some packets of delay $d_i^\star$ that is served ($0 \leq \tilde{B}_{i,d_i^\star} < B_{i,d_i^\star}$). The amount is decided by the remaining power among the serviced users.

- After power allocation and the corresponding packet service, θ is decided by the remaining traffic, given as

$$\theta = \min\left\{1, \frac{1}{\sum A_i}\left[\Psi - \sum_{i=1}^{M}\sum_{\tau=2}^{d_i^\star} \tau \cdot B_{i,\tau}\right]\right\}. \tag{5.145}$$

- The total power constraint is satisfied by

$$\sum_{i=1}^{M} P_i^\star \leq P_{total}. \tag{5.146}$$

5.6 Simulation Results

We now compare the simulation performance of the algorithm with the steady state analytic results developed in Section 5.4 for the phased array antenna and the multiple beam antenna. We consider 49 ($= M$) users uniformly located in a 7-by-7 planar grid.

The satellite is assumed to be able to provide up to 20 ($= K$) signals simultaneously. The user traffic is assumed to be linearly distributed with an arrival rate $A_i = i \cdot \beta$ for $i = 1, \cdots, M$, where β is a constant slope. In the simulation, two forms of real-time traffic are considered with the same arrival rate A_i: Poisson arrival random traffic and constant streaming deterministic traffic. We change the distance l between neighboring users from $0.01\frac{\lambda L}{D}$ to $\frac{\lambda L}{D}$. With the Poisson arrival traffic, we perform 5 simulations with 500 iterations for each and average them for every distance. The streaming traffic gives an identical result for every simulation due to its deterministic traffic pattern.

For a steady-state analysis, we consider an average interference level, so that the throughput decreases monotonically until only a single active signal is used. In Section 5.4, we compared the steady-state performances of the multiple beam antenna and the phased array antenna, based on the analytic closed-form results, and assumed the perfect scheduling if the total coverage area of K orthogonal beams is smaller than the total area occupied by M users. This does not hold any more here for a more accurate comparison with the algorithm simulation.

Fig. 5-12 plots the time average of $\theta(t)$ from simulation results for the two types of traffic and compares them with steady-state analytic solutions. In this example, the algorithm achieves 97.5 % of the steady-state performance with the stream traffic and 94.1 % with the Poisson traffic, in case of no interference at $l = \frac{\lambda L}{D}$. The algorithm is simulated in discrete timeslots while the steady-state analysis is assumed to use idealistic scheduling in continuous time. Thus, the simulation may not use all K signals even with no interference (which is true for the Poisson arrival traffic of this example and shown in Fig. 5-13). When a small number of demanding users consume more or whole onboard power, other users cannot be served at the same time, reducing the number of active signals, and thus the throughput. In addition, the random and bursty Poisson arrival traffic deviates from the steady-state traffic pattern. As a result, the number of active signals used and the throughput are reduced further, compared to the constant stream traffic that is a good approximation to the steady-

state. This example shows that efficient resource scheduling for random bursty data traffic is more difficult than scheduling for steady circuit traffic.

In the middle range of distance, $0.25\frac{\lambda L}{D} < l < 0.6\frac{\lambda L}{D}$, the simulation performance of the Poisson arrival traffic is not so close to the steady-state solution and the performance of the stream traffic as in the other range of distance. This is due to the sub-optimum selection process, which is simplified to update the interference level only after user selection and can degrade the performance in the middle range where the better choice between interference suppression and sequential service can change frequently. Nevertheless, the algorithm still achieves more than 85% of the steady-state solution except for a small range of distances. The steep decrease of the average throughput of simulations at $l = 0.25\frac{\lambda L}{D}$ is due to the distance threshold value $l^* = 0.25\frac{\lambda L}{D}$ in this simulation, which is the optimum value chosen in a heuristic manner. Users at the distance of $l < l^*$ from already selected active users are not selected though they have higher values of $\alpha_i^2 H_i d_i^{max}$ than others because their selection will decrease the capacities of some of already selected users significantly. (The worse performance of interference suppression than that of sequential service in Fig. 5-4 and 5-5 is reminded at very small distance between active users.) The simulation performance depends heavily on the parameter of l^*. The distance threshold l^* decides how many active signals are used simultaneously under interference. If the value is too big, the algorithm under-performs for small interference because it averts interference too much and loses the throughput gain from having multiple parallel channels. If the value is too small, the algorithm may send too many signals under severe interference and lose efficiency.

On the other hand, in addition to continuous scheduling, the steady-state analysis assumes the use of all K active signals with the perfect selection process of K active users all the time if the total coverage area of K orthogonal beams is smaller than the total area occupied by M users, and thus gives an upper-bound to any sub-optimum scheduling algorithm. In this example, the deterministic stream traffic uses all K signals near $l = \frac{\lambda L}{D}$ and achieves near-optimum performance for every distance except

at $l < 0.25\frac{\lambda L}{D}$. The multiple beam antenna shows a significant loss of the throughput as the distance decreases and the interference level increases, because interference suppression cannot be deployed and the only option is to reduce the number of active beams (with a fixed amount of transmission power for each beam).

Fig. 5-13 shows close correlation between the average throughput (the left axis) and the average number of active signals used (the right axis) for the Poisson arrival traffic. Thus, the key point to the optimum scheduling is to increase the number of active signals unless interference becomes too severe with the optimum selection of the value of l^*.

5.7 Summary

With the use of SSPA, a phased array antenna can generate flexible dynamic (≤ 1 msec) multiple signals and is appropriate for the agile beam satellite system. The optimum design of a phase array antenna transmission system in a large scale can improve the efficiency of satellite resource allocation, especially for bursty data traffic. In this chapter, we have found the solution for joint antenna gain patterning and scheduling for the phased array antenna. The optimum scheme is to provide the narrowest spotbeams for the non-interfering active users that are located far enough. When interference is significant between close-in users, the optimum pattern, which depends on the distances between users and the SNR, can be one of the following: complete cancellation of interference, optimum suppression of interference, and the sequential service of close-in users. Subject to average delay constraints, signals should be switched and gain-patterned according to channel conditions, interference levels, and the marginal returns of the composite cost function with respect to allocated capacity. Due to flexible power allocation, the phased array antenna can provide better performance than the multiple beam antenna when a small number of users are very demanding or many users are densely crowded in a small area.

Then, we developed a near-optimum and simple real-time algorithm for user selec-

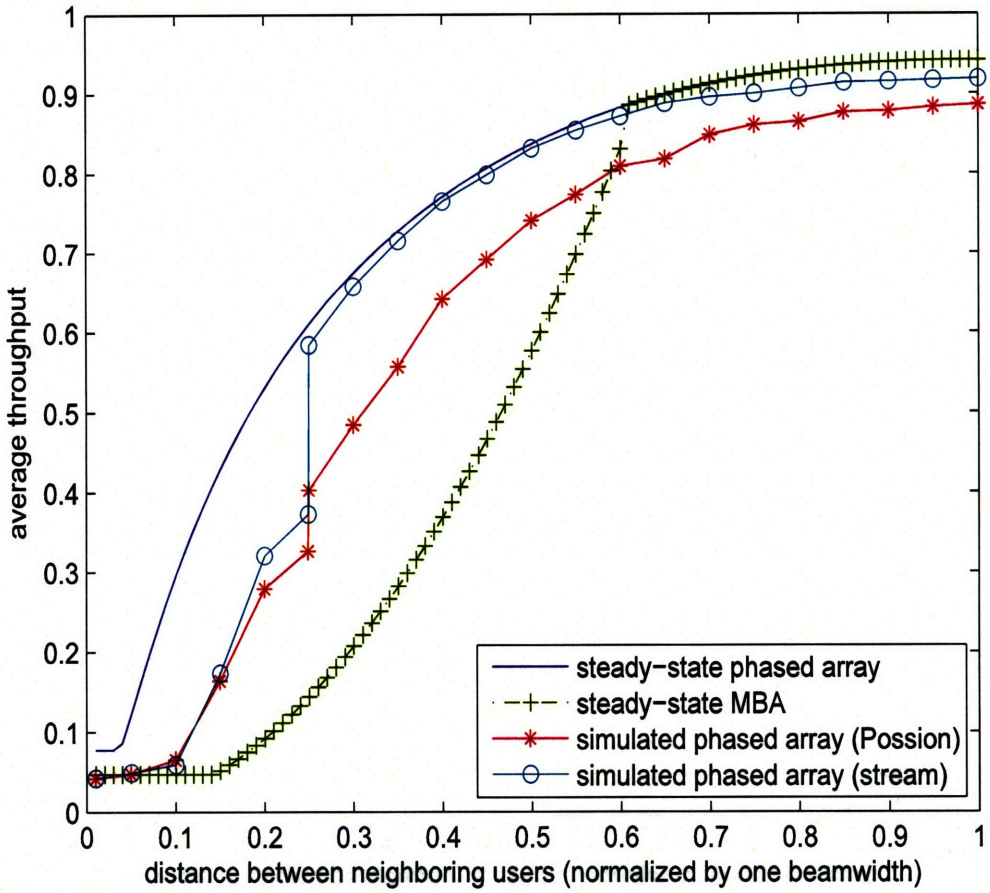

Figure 5-12: Comparison of average throughputs between algorithm simulation (with Poisson arrival random traffic and constant stream) and steady-state analytic solutions (for the phased array antenna and the multiple beam antenna)

tion, antenna gain patterning, power allocation and admission control. The algorithm serves users with better channel conditions, less interference and higher queuing delays. Power is allocated for the selected users to have the same marginal returns of a cost function, which depends on throughput, delay and channel conditions, with respect to consumed power. We introduced a total accumulated delay constraint for admission control. The simulation result showed that the algorithm can achieve up to 94% of the steady-state analytic result for random traffic.

Here we considered interference suppression and interference-free sequential ser-

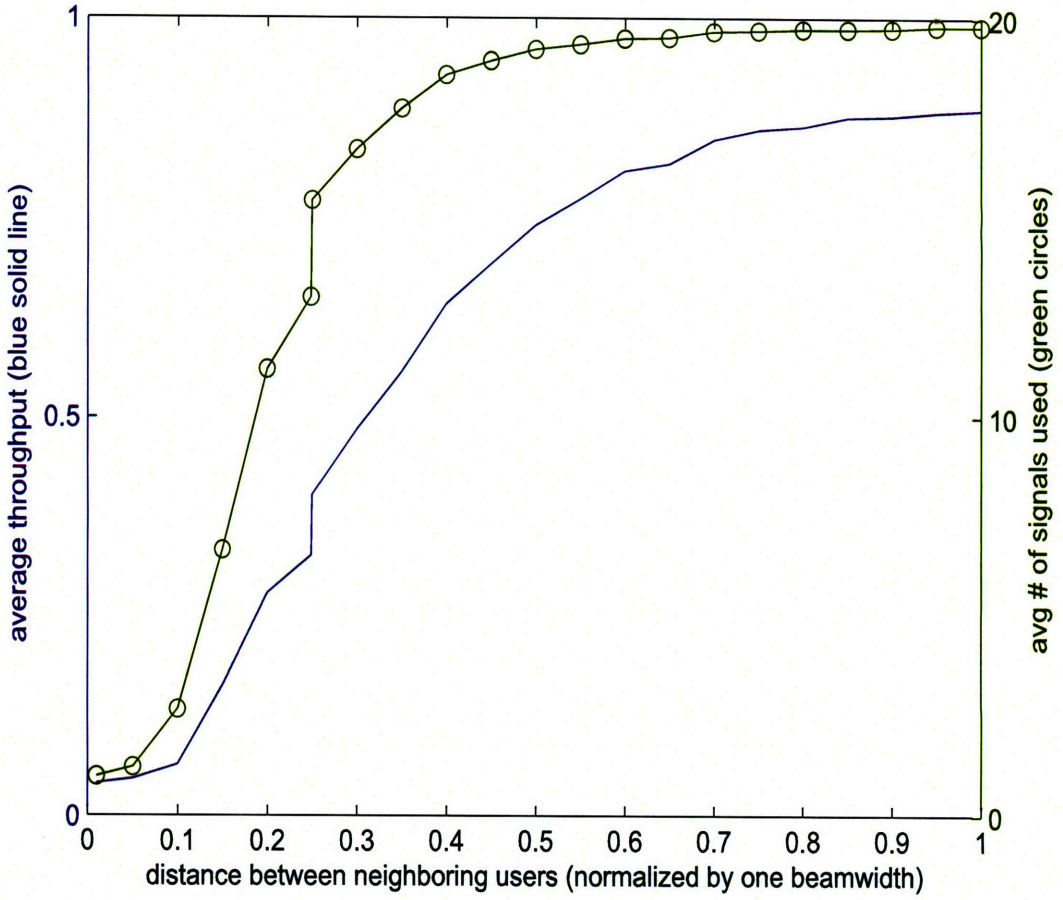

Figure 5-13: A plot of the average throughput (the left axis) and the average number of active signals (the right axis) for the Poisson arrival traffic

vice in the viewpoint of the satellite transmitter. Interbeam interference degrades system performance and should be avoided if possible. In multibeam and multi-user environments, a receiver equipped with an array antenna and array signal processing can also suppress interference and isolate the desired signal [49]. Intentional interference from an outside jammer can be suppressed only by the receiver beamforming. A smart antenna can be installed in the user terminal.[6] It estimates the direction of arrival of the desired signal and calculates beamforming vectors that suppress inter-

[6]Whether a smart antenna can be installed in the user terminal depends on the operating frequency band due to the minimum antenna length constraint. The frequency band should be high enough for the smart antenna to fit in a small satellite handheld phone. Current systems such as Iridium and Globalstar operate too low carrier frequencies to deploy smart antennas.

ference from other signals. By exploiting transmission beamforming and scheduling as well as array signal processing in receivers, one can improve signal quality and increase system capacity significantly. Since the satellite has all the information or good estimates of traffic distribution and channel conditions, transmission beamforming is straightforward to deploy with onboard signal processing.

5.8 Appendix: Complex Gradients in [54]

Here we present a convenient method to solve a real-valued optimization problem that has a complex vector $\mathbf{V} = \mathbf{V_x} + j\mathbf{V_y}$. This approach has been developed in Brandwood's work [6] and summarized in Van Trees's textbook [54].

First, we consider a complex scalar variable $V = V_x + jV_y$ and a function of interest $f(V) = f(V_x, V_y)$. We define a function of

$$g(V, V^*) \equiv f(V_x, V_y), \tag{5.147}$$

which is analytic with respect to z and z^* independently. It can be shown that

$$\frac{\partial g(V, V^*)}{\partial V} = \frac{1}{2}\left(\frac{\partial f(V_x, V_y)}{\partial V_x} - j\frac{\partial f(V_x, V_y)}{\partial V_y}\right) \tag{5.148}$$

and

$$\frac{\partial g(V, V^*)}{\partial V^*} = \frac{1}{2}\left(\frac{\partial f(V_x, V_y)}{\partial V_x} + j\frac{\partial f(V_x, V_y)}{\partial V_y}\right) \tag{5.149}$$

Then, a necessary and sufficient condition for $f(V)$ to have a stationary point is one of the following two:

$$\frac{\partial g(V, V^*)}{\partial V} = 0, \tag{5.150}$$

where V^* is treated as a constant in the partial differentiation, or

$$\frac{\partial g(V, V^*)}{\partial V^*} = 0, \tag{5.151}$$

where V is treated as a constant in the partial differentiation. Eq. (5.151) has been used in Section 5.2.

For a vector $\mathbf{V}$, the complex gradient operator is defined as

$$\nabla_{\mathbf{V}} = \left[\frac{\partial}{\partial V_1}, \cdots, \frac{\partial}{\partial V_M}\right], \tag{5.152}$$

where

$$\frac{\partial}{\partial V_i} \equiv \frac{\partial}{\partial V_{i,x}} - j\frac{\partial}{\partial V_{i,y}}. \tag{5.153}$$

For a conjugate transpose $\mathbf{V^H}$, we define

$$\nabla_{\mathbf{V^H}} = \left[\frac{\partial}{\partial V_1^*}, \cdots, \frac{\partial}{\partial V_M^*}\right], \tag{5.154}$$

where

$$\frac{\partial}{\partial V_i^*} \equiv \frac{\partial}{\partial V_{i,x}^*} + j\frac{\partial}{\partial V_{i,y}^*}. \tag{5.155}$$

Then, for a real-valued function of $f(\mathbf{V}) = f(\mathbf{V_x}, \mathbf{V_y}) \equiv g(\mathbf{V}, \mathbf{V^H})$, which is analytic with respect to $\mathbf{V}$ and $\mathbf{V^H}$ independently, a necessary and sufficient condition in which $f(\mathbf{V})$ has a stationary point is either

$$\nabla_{\mathbf{V}} g(\mathbf{V}, \mathbf{V^H}) = 0, \tag{5.156}$$

where $\mathbf{V^H}$ is treated as a constant, or

$$\nabla_{\mathbf{V^H}} g(\mathbf{V}, \mathbf{V^H}) = 0, \tag{5.157}$$

where $\mathbf{V}$ is treated as a constant.

Chapter 6

Conclusions

6.1 Summary

As frequency bands go up for high data rate broadband communications, future satellites will be able to synthesize narrow spotbeams to project high power density. Multiple active beams give better throughput than a single beam by frequency reuse and by virtue of the concavity of the capacity function with respect to power. However, due to high cost and heavy weight of carrying many active beams and transponders to service the whole coverage area simultaneously, it is only feasible to time-share a small number of modulators and transmitter power. In this thesis, we solved idealized joint optimization problems of resource allocation/scheduling, congestion control and antenna gain patterning for satellite communications based on traffic demand and channel conditions, to obtain theoretical steady-state bounds and develop an optimum-approaching on-line algorithm.

We modeled two types of transmission antennas: a multiple beam antenna equipped with traveling wave tube amplifiers (TWTA) and a phased array antenna with solid state power amplifiers (SSPA). Every multiple beam antenna feed is fed by a single TWTA, which is driven up to saturation for efficiency, and constraints the power for each beam to the maximum power of a single TWTA. On the other hand, since an antenna-patterning matrix and SSPAs can superimpose signals linearly at array

elements, the phased array antenna can allocate signal power to a single user up to the total power of the array. With advancement of electronic and electro-optical switching technologies, the phased array antenna and SSPAs can provide advantages of better linearity, more flexible beam shape/size and faster scheduling/cycling than the multiple beam antenna and TWTAs. Our analysis indicates that the phased array antenna can provide superior performance when a small number users are very demanding or there are many users in a small area (compared to a diffraction-limited beam size).

For power and beam allocation, we suggested and compared different cost functions, in order to gain insights on the trade-offs between maximum total capacity and proportional fairness. Substantial power gains and fairness advantages can be realized by using optimum power allocation for parallel multibeams. We considered a realistic situation, where the numbers of active beams, TWTAs and the corresponding modulators are less than that of the cells in the coverage area, and showed that a modest number of active parallel beams are sufficient to cover many cells efficiently.

We developed a jointly optimized scheme of resource allocation and congestion control with transmitter-sharing and average delay constraints for a multiple beam antenna system. Congestion control is required to prevent excessive packet loss and stabilize the system within an acceptable queueing delay. We modeled admission control as a simple back-off parameter that is fed to the users and throttles incoming traffic rate. The jointly optimum scheme gives the system throughput in terms of (i) the average of expected values of service cell parameters, such as incoming traffic rate, transmission rate and average delay deadline, or (ii) the parameters of the most dominant cell. Numerical examples showed that this scheme for the multiple beam antenna outperforms uniform beam allocation by providing higher throughput (e.g., a factor of 2 in the case of linearly distributed traffic across the cells) and/or smaller average queueing delays.

We found the solution for joint antenna gain patterning and scheduling by considering spatially close co-channel interference in the use of phased array antenna.

When users are located far enough, the optimum scheme is to provide the narrowest spotbeams for the non-interfering active users. When potential interference can be significant between close-in users, the optimum pattern, which depends on the distances between users and the signal-to-noise ratio (SNR), can be (i) complete cancellation of interference, (ii) optimum suppression of interference or (iii) sequential service. We suggested an optimum scheduling policy, which selects users with higher marginal returns of a composite cost function with respect to allocated power, in terms of better channel conditions, less interference (depending on users' geographic distribution), and larger delay. From the optimum analytic result, we suggested a near-optimum real-time algorithm of performing active user selection, power allocation, antenna gain patterning and admission control. The simulation result showed that the algorithm can achieve a throughput close to the analytic steady-state upper bound (up to 94% of the steady-state solution with random traffic).

6.2 Comments

Commercial satellites have expanded applications from traditional telephone trunking and TV broadcasting to mobile voice phones and Internet access services. A huge success of satellite digital radios and a gradual increase of satellite Internet services suggest that there can be a huge potential market for new applications of communication satellites. For multimedia and other data services over satellite networks, the efficient management of satellite communication resources is crucial for the economic competitiveness of the medium. Since satellite on-board resources are expensive, an optimized scheme of the agile antenna gain pattern and beam scheduling can improve the efficiency of transmission and power management further.

In this thesis, we only considered the transmission scheduling problem over satellite-to-Earth downlinks. Uplink transmission can also benefit from scheduling algorithms based on traffic demand and channel conditions. The difference is that uplink accesses occur from many users on the Earth to a single satellite in a distributed manner,

which requires more rigorous coordination of random user transmission and power control. Otherwise, multiple access interference (MAI) and a significant loss in system throughput will occur.

A similar scheme of resource allocation and scheduling can be applied to terrestrial wireless communications, for which different channel characteristics should be addressed. The wireless channel shows more scattering in the order of $r^{-4} \sim r^{-6}$ compared to r^{-2} of the free space loss over the satellite channel, where r is the distance between a transmitter and a receiver. Multipath fading is a dominant factor for fast and deep fading events. The wireless channel also has a low chance of clear line-of-sight (LOS) signals due to blocking by buildings, trees, etc.

The long propagation delay over satellite channels makes it hard to monitor the system status and environments perfectly and to adjust system parameters optimally. The lack of perfect information on channels and traffic distribution prevents instantaneous and optimal control of traffic arrival rates. However, due to slow fading and good line-of-sight signals over satellite channels, the satellite system is relatively stable and quasi-static adaptation can yield substantial gains. The system can take advantage of active inference and estimation of the information based on past measurements of the data channel, estimates on a reciprocal channel or direct feedback from a return channel. Still, the development of optimum interactive protocol (e.g., TCP) over a high-latency satellite channel is a challenging task and remains a big part of future work.

This thesis focused on optimizing single-satellite transmission. Considering a network of multiple satellites requires solving additional problems of handover and inter-satellite link routing. Our scheme of resource scheduling will be applicable to each satellite in the network in a distributed way.

Appendix A

Notation

We provide a list of symbols used in the thesis. Indices for user, cell and time are suppressed for applicable symbols. For example, A is listed instead of A_i.

$\bar{\cdot}$	time average operator (overbar)
A	average arrival rate of incoming traffic
$A_{coverage}$	satellite coverage area
B	amount of traffic having waited for some delay in queue
C	channel capacity
D	transmit antenna size
$E[\cdot]$	ensemble average operator
E_b	average signal energy per bit
F	amount of accumulated traffic demand
$F_{coverage}$	footprint diameter of total satellite coverage area
$\mathcal{F}$	two-dimensional Fourier transform
G	amplitude of aperture distribution for phased array antenna
H	power loss by deploying interference suppression
J	Lagrangian function
K	number of active beams
L	satellite altitude

M	number of users
M_{act}	number of active users
N	number of cells
N_0	white noise power density
P	amount of allocated power
P_0	maximum power for each beam of multiple beam antenna
P_{total}	maximum total power of phased array antenna
$\Pr[\cdot]$	probability of event
Q	average power allocated when signal is on for phased array antenna
R	average transmission rate
S	sum of arrival rates for every cell except the highest
$SINR$	signal-to-noise-and-interference ratio
T	time interval
U	received signal on Earth for phased array antenna
V	aperture field distribution for phased array antenna
W	bandwidth
X	dummy variable and some constant
Y	dummy variable and some constant
Z	additive white Gaussian noise in interference channel
a	scaling constant for proportional fairness
c	scaling constant in interference channel
d	queueing delay of packet
e	packet error rate
f	several functions including utility function in Chapter 5
g	power gain function of parallel multibeams with optimum power allocation over uniform allocation
h	user index
i	user/cell index
j	user/cell index

k	user index
l	distance between adjacent users
$\bar{l}_p$	average packet size
m	user index
n	order of cost function
p	probability density function (PDF)
q_{max}	maximum queue size
s	desired signal component of phased array antenna
t	time index
u	utilization factor
v	time varying waveform
x	Cartesian coordinate on Earth and dummy variable
y	Cartesian coordinate on Earth
z	binary indicator for beam allocation
Γ	threshold function of 1st order cost function
Δ	average delay deadline
Λ	Lagrangian multiplier
Φ	unconsumed power in Appendix of Chapter 3
Ψ	upper bound to total of accumulated delay
Ω	set of active users
α	signal attenuation due to weather condition
β	parametric slope of linearly distributed traffic
γ	coefficient of complete cancellation scheme
δ	receiver antenna size
ϵ	difference between accumulated traffic and allocated capacity
ζ	total service time during time interval for multiple beam antenna
η	Cartesian coordinate on antenna aperture
θ	congestion control parameter
κ	Lagrangian multiplier

λ	carrier wavelength
μ	Lagrangian multiplier and normalizing factor of optimum suppression
ν	Lagrangian multiplier
ξ	Cartesian coordinate on antenna aperture
ρ_0	maximum power density of phased array antenna element
σ^2	variance of amount of accumulated traffic
τ	time index
ϕ	parametric slope of linearly distributed channel capacity
ψ	coefficient of optimum suppression scheme
ω	phase of aperture distribution for phased array antenna

Bibliography

[1] I. F. Akyildiz, G. Morabito and S. Palazzo, "TCP-Peach: A New Congestion Control Scheme for Satellite IP Networks," *IEEE/ACM Trans. on Networking*, vol. 9, no. 3, pp. 307-321, June 2001.

[2] G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, 5th ed., Academic Press, 2000

[3] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1995.

[4] D. P. Bertsekas and R. G. Gallager, *Data Networks*, 2nd ed., Prentice Hall, 1992.

[5] C. J. Black, P. Takats, M. Cote and T. T. Le-Ngoc, "Data Communication Satellite System and Method of Carrying Multimedia Traffic," US Patent No. 6,377,561, Spar Aerospace Ltd., Mississauga, CA, April 2002.

[6] D. H. Brandwood, "A Complex Gradient Operator and Its Application in Adaptive Array Theory," *Proc. IEE*, Special Issue on Adaptive Arrays, vol. 130, pp. 11-17, February 1983.

[7] V. W. S. Chan, 6.976 Lecture Notes for Space Communications and Networks, MIT EECS, 2002.

[8] J. P. Choi, "Channel Prediction and Adaptation over Satellite Channels with Weather-Induced Impairments," Master's Thesis, EECS, MIT, May 2000.

[9] J. P. Choi and V. W. S. Chan, "Predicting and Adapting Satellite Channels with Weather-Induced Impairments," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 779-790, July 2002.

[10] J. P. Choi and V. W. S. Chan, "Optimum Multibeam Satellite Downlink Power Allocation Based on Traffic Demands," IEEE Globecom 2002, Taipei, Taiwan.

[11] A. K. Clarke, "Peacetime Uses for V2," *Wireless World*, p. 58, February 1945. (text available online at http://lakdiva.org/clarke/1945ww/1945ww_feb_058.html)

[12] M. H. M. Costa, "On the Gaussian Interference Channel," *IEEE Trans. on Information Theory*, vol. IT-31, no. 5, pp. 607-615, September 1985.

[13] T. M. Cover, "Broadcast Channels," *IEEE Trans. on Information Theory*, vol. IT-18, no. 1, pp. 2-14, January 1972.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[15] L. L. Dai, J. P. Choi and V. W. S. Chan, "Satellite and Space Communications - Technologies and Systems," to appear in *Encyclopedia of Life Support Systems (EOLSS)*, UNESCO, 2006.

[16] S. Egami, "A Power-Sharing Multiple-Beam Mobile Satellite in Ka Band," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 2, pp. 145-152, February 1999.

[17] S. Egami and M. Kawai, "An Adaptive Multiple Beam System Concept," *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, no. 4, pp. 630-636, May 1987.

[18] Eutelsat website, http://www.eutelsat.com.

[19] P. M. Freitag and S. R. Forrest, "A Coherent Optically Controlled Phased Array Antenna System," *IEEE Microwave and Guided Wave Letters*, vol. 3, no. 9, pp. 293-295, September 1993.

[20] A. Fu, E. Modiano and J. Tsitsiklis, "Optimal Energy Allocation for Delay-Constrained Data Transmission over a Time-Varying Channel," IEEE Infocom 2003, San Francisco, April 2003.

[21] A. Fu, E. Modiano and J. Tsitsiklis, "Optimal Energy Allocation and Admission Control for Communications Satellites," *IEEE/ACM Trans. on Networking*, vol. 11, no. 3, pp. 488-500, June 2003.

[22] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, 1968.

[23] A. Ganti, "Transmission Scheduling for Wireless and Satellite Systems," PhD Thesis, MIT EECS, February 2003.

[24] Gilat website, http://www.gilat.com.

[25] Globalstar website, http://www.globalstar.com/en

[26] J. Goodman, *Introduction to Fourier Optics*, 3rd ed., Roberts & Company, 2005.

[27] T. S. Han and K. Kobayashi, "A New Achievable Rate Region for the Interference Channel," *IEEE Trans. on Information Theory,* vol. IT-27, no. 1, pp. 49-60, January 1981.

[28] S. Haykin, *Communication Systems*, 3rd. ed., Wiley, 2004.

[29] T. R. Henderson and R. H. Katz, "Transport Protocols for Internet-Compatible Satellite Networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 2, pp. 326-344, February 1999.

[30] Inmarsat website, http://www.inmarsat.com.

[31] Intelsat website, http://www.intelsat.com.

[32] A. Jamalipour and A. Ogawa, "Packet Admission Control in a Direct-Sequence Spread-Spectrum LEO Satellite Communications Network," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1649-1656, October 1997.

[33] R. C. Johnson, *Antenna Engineering Handbook*, 3rd ed., McGraw-Hill, 1993.

[34] D. Katabi, M. Handley and C. Rohrs, "Internet Congestion Control for High Bandwidth-Delay Product Networks," ACM Sigcomm 2002, Pittsburgh, August 2002.

[35] F. P. Kelly, A. K. Maulloo and D. K. H. Tan, "Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability," *Journal of the Operational Research Society*, vol. 49, pp. 237-252, 1998.

[36] P. Khan, L. Epp and A. Silva, "A Ka-Band Wideband-Gap Solid-State Power Amplifier: Architecture Identification," IPN Progress Report 42-162, Jet Propulsion Laboratory, Pasadena, CA, August 15, 2005.

[37] K.-T. Ko and B. R. Davis, "A Space-Division Multiple-Access Protocol for Spot-Beam Antenna and Satellite-Switched Communication Network," *IEEE Journal on Selected Areas in Communications*, vol. SAC-1, no. 1, pp. 126-132, January 1983.

[38] H. Koraitim and S. Tohme, "Resource Allocation and Connection Admission Control in Satellite Networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 2, pp. 360-372, February 1999.

[39] E. Lutz, M. Werner and A. Jahn, *Satellite Systems for Personal and Broadband Communications*, Springer, 2000.

[40] G. Maral and M. Bousquet, *Satellite Communications Systems: Systems, Technoques and Technology*, 4th ed., John Wiely & Sons, 2002.

[41] D. Martin, *Communication Satellites 1958-1995*, The Aerospace Corporation, 1996.

[42] P. F. McManamon, T. A. Dorschner, D. L. Corkum, L. J. Friedman, D. S. Hobbs, M. Holz, S. Liberman, H. Q. Nguyen, D. P. Resler, R. C. Sharp and E. A. Watson, "Optical Phased Array Technology," *Proceedings of the IEEE*, vol. 84, no. 2, pp. 268-298, February 1996.

[43] U. K. Mishra, P. Parikh and Y.-F. Wu, "AlGaN/GaN HEMTs - An Overview of Device Operation and Applications," *Proceedings of the IEEE*, vol. 90, no. 6, pp. 1022-1031, June 2002.

[44] M. J. Neely, E. Modiano and C. E. Rohrs, "Power Allocation and Routing in Multibeam Satellites with Time-Varying Channels," *IEEE/ACM Trans. on Networking*, vol. 11, no. 1, pp. 138-152, February 2003.

[45] M. J. Neely and E. Modiano, "Fairness and Optimam Stochastic Control for Heterogeneous Networks," IEEE Infocom 2005, San Francisco, CA.

[46] D. Pozar, *Microwave Engineering*, Addison-Wesley, 1993.

[47] J. G. Proakis, *Digital Communications*, 3rd ed., McGraw-Hill, 1995.

[48] J. Romero-Garcia and R. De Gaudenzi, "On Antenna Design and Capacity Analysis for the Forward Link of a Multiple Power Controlled Satellite CDMA Network," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 7, pp. 1230-1244, July 2000.

[49] P. Saengudomlert and V. W. S. Chan, "Hybrid Optical and Electronic Signal Processing for Ultra-Wideband RF Antenna Arrays," IEEE ICC 2005, Seoul, Korea.

[50] H. Sato, "On the Capacity Region of a Discrete Two-User Channel for Strong Interference," *IEEE Trans. on Information Theory*, vol. IT-24, no. 3, pp. 377-379, May 1978.

[51] J. Siwko and I. Rubin, "Connection Admission Control for Capacity-Varying Networks with Stochastic Capacity Change Times," *IEEE/ACM Trans. on Networking*, vol. 9, no. 3, pp. 351-360, June 2001.

[52] D. N. C. Tse and S. V. Hanly, "Linear Multiuser Receivers: Effective Interference, Effective Bandwidth and User Capacity," *IEEE Trans. on Information Theory*, vol. 45, no. 2, pp. 641 - 657, March 1999.

[53] D. N. C. Tse and O. Zeitouni, "Linear Multiuser Receivers in Random Environments," *IEEE Trans. on Information Theory*, vol. 46, no. 1, pp. 171 - 188, January 2000.

[54] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation and Modulation Theory*, Wiley-Interscience, 2002.

[55] S. Wilson, *Digital Modulation and Coding*, Prentice Hall, 1995.

[56] W. Yang and G. Xu, "Optimal Downlink Power Assignment for Smart Antenna Systems," IEEE ICASSP 1998, Seattle, WA.